Developing Methods for Reaction Informatics and Automation

by

Rui Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in the University of Michigan
2024

Doctoral Committee:

Assistant Professor Tim Cernak, Chair
Professor John Montgomery
Assistant Professor Wenhao Sun
Professor Paul Zimmerman

Rui Zhang

rzqcr@umich.edu

ORCID iD: 0000-0001-7396-7165

Acknowledgements

First and foremost, I would like to thank Dr. Tim Cernak for his guidance and support throughout the work going into this thesis, and supplying our lab with stream of both efficient instruments and novel ideas. I am honored to be one of the early members of the Cernak lab, getting as close to a ground-floor experience as possible with the setup of new instruments, submission of the first manuscript, reviewing of papers and grants, and much more.

I am also grateful towards my committee members, Dr. Paul Zimmerman, Dr. John Montgomery, Dr. Wenhao Sun, and previous member Dr. Aaron Frank for their valuable comments about various project progress and directions.

The members of the Cernak Lab, each equipped with diverse skillsets, are a valuable source of knowledge and expertise, some of which are featured in this thesis. In particular, I would like to thank, in order of appearance in this thesis:

Dr. Babak Mahjour for performing the preliminary reaction enumeration and retrosynthetic analysis work in Chapter 2, and whose code formed half of my python education.

Dr. Clint Regan for the other python half, and also providing valuable guidance transitioning into matrix methods for reaction enumeration.

Andrew Outlaw, though not directly featured, for experimental work in novel amine–acid couplings that bolster the accompanying manuscripts by demonstrating reaction space expansion in an experimental setting.

Dr. Yingfu Lin for the immense amount of synthetic work in the various routes towards stemoamide, and adapting key computer-proposed steps into experiment, affirming that professional expertise is still indispensable in the age of computers.

Di Wang, for preliminary work on the concept of matrix-encoded synthetic intermediates and the notion of common substructure size as a distance metric.

Andrew McGrath and honorary member Khadija Shafiq for experimental work in optimizing conditions for the esterification reaction featured in Chapter 4.

Dr. Tim Hopper, Dr. Brian Kelly, and Dr. Pat Walters of Relay Therapeutics are thanked for graciously performing docking studies featured in Chapter 2.7.

Though I do not regularly perform experimental work, as the lab ordering representative I am grateful for members of the procurement and loading dock staff – Carrie Vaquera, Stasi Mehaffey, Hawaii Maliga, Chris Bluteau, Ron Farnstrom, and surely more up the administrative ladder – for keeping chemicals and labware moving behind the scenes.

Table of Contents

List of Figures

List of Appendices

Abstract

As the field of organic synthesis enters the digital age, an unprecedented number of tools are opened up for analyzing, planning and executing chemical reactions. This thesis will describe the development of two methods for analyzing reactions between organic molecules, and one for the automation of setting up high-throughput experiments.

First, an extensive exploration of amine–carboxylic acid reaction space was conducted through computational enumeration of all theoretically possible matrix-encoded transformations between a simple amine–acid pair, delving into an under-explored axis of chemical space exploration. The extent of physicochemical property modulation enabled by this technique is analyzed using both small and large building blocks. The performance of reaction enumeration method in generating virtual libraries from one single building block pair was evaluated against the conventional approach of coupling many building blocks through one robust reaction.

Next, the matrix encoding technique was applied to analysis of reactions throughout a total synthesis route. A new method of synthetic route visualization was developed by charting the graph edit distance between each intermediate and the target, producing a graph from which valuable high-impact steps can be quickly identified and analyzed. By merging this technique with computer-aided synthesis planning software, two enantioselective syntheses of the alkaloid stemoamide were conducted. At the length of six and three steps respectively, they mark the shortest synthesis of this molecule to date.

Lastly, a platform for automated setup of high-throughput experiments in 24- and 96-wellplates was developed, combining the online electronic notebook phactor™ with the Opentrons OT-2 autopipettor. Benchmarking test against the existing manual workflow reveal that the robotic setup performs adequately in most conditions, with the main exception being poorly soluble solid reagents. An alternative manual dosing of such solids was developed using custom spatula designs. The automated platform enables many novel experimental designs, such as remote collaboration over teleconferencing software, and small-scale library synthesis of nearly 100 products in a single wellplate.

Chapter 1 Introduction and Manuscript Overview

1.1 Background of data science and automation in the field of organic chemistry

The presence of data informatics in organic chemistry predates the incorporation of computers in the field. Methods of enumerating alkanes were proposed in 1875 by Cayley[1] and iterated upon by many others[2–4]. Automated synthesis of peptides was developed as early as 1966 by Merrifield and Jernberg[5]. On the computation front, Corey and Wipke reported in 1969 a system that allowed conversion of molecular structures via a drawing pad to a machine-readable and displayable format[6]. These structures could then be analyzed and broken down *in-silico* through encoded reaction rules, generating trees of suggested retrosynthetic disconnections.

As improvements are made in computational power, information storage, and data science techniques, they have readily been harnessed in furthering the field of organic synthesis. These developments offer abundant opportunities to either add novel methods to this common toolbox, or tailor existing tools to specific laboratory tasks. In this manuscript, three contributions in the area of computers in chemistry will be detailed.

The exploration of chemical space through enumeration was first performed as a mathematical exercise by Cayley[1] and Schiff[3], both developing methods to calculate the number of acyclic alkanes for a given carbon count. Many subsequent publications either offered corrections to the proposed formulae[2,4], or expanded their applications to other functionalities[7,8]. In contrast, modern enumeration of chemical structures aim to produce a virtual collection of molecules that span a given chemical space. For example, the GDB-17 database by Reymond[9] contains structures that have a maximum of 17 heavy atoms, only possess a subset of C, N, O, S, and halogen atoms, and filtered to remove highly strained structures and functional groups. Pharmaceutical companies and chemical suppliers provide databases of molecules that can be quickly synthesized by coupling on-hand building blocks with tried-and-true reactions. Examples of these databases include the Proximal Lilly Collection[10], Pfizer Global Virtual Library[11] and Enamine REAL library[12].

Another type of enumeration is that of chemical reactions. Systematic methods for classifying organic reactions have been proposed by researchers such as Hendrickson[13], intended to aid reaction cataloging and referencing, while numerical methods developed by Ugi and coworkers[14,15], that encoded chemical reactions as changes in a matrix of bond orders, were employed to discover novel reactions by fixing the pattern of bond changes while iterating through identities of the atoms at which these transformations occur[16].

As a complement to previous works, this manuscript will conduct a third axis of enumeration. Using matrix-encoded structures and transformations, the coupling space between a single pair of amine and carboxylic building blocks, chosen for their abundance and comparative low cost, will be extensively explored. Modulations of physicochemical properties will be compared between simple small molecules as well as larger druglike molecules, and the performance of this method in generating ultra-large virtual libraries will be evaluated.

Moving forward from exploring single-step reactions, this manuscript will next apply matrix methods to evaluating multi-step synthetic routes. It is notable that several early works in representing molecules as matrix-encoded graphs, with atoms as nodes and bonds as edges, have been geared towards developments for retrosynthetic software[15,17], and this encoding method remains in use among contemporary reaction prediction literature[18–20].

When converting computer-proposed retrosynthesis into experimental routes, methods of assessing step impact are useful to locate high-impact steps to ideally preserve in the experimental process. Existing assessment methods fall between two main frameworks. One classifies each step under one or several reaction types, with which to evaluate their synthetic ideality. For example, reactions that form rings[17], merge multiple building blocks into one[21,22], form multiple bonds in a cascading mechanism[23], or C–C bonds in strategic locations[24] are considered highly impactful. In contrast, synthetic steps that manipulate protecting groups[25], undergo unnecessary redox operations[26], or perform multiple functional group interconversions[27] are deemed to have low or negative impact. Evaluations of step impact using reaction type are simple for chemists to understand in context of their existing knowledge, but the method does not intrinsically differentiate two steps that fall under the same reaction class. The second assessment method calculates given properties for each intermediate, most commonly a measure of molecular complexity[28–30] but also physicochemical properties such as molecular weight or fraction of sp^3 atoms[31]. A graph of property against intermediate readily shows the impact of each step as their slopes, where steps with the highest slope value are assigned highest impact. However, complexity of algorithms employed to compute these properties result, to varying extents, in a "black box" visualization – the graphs can show which steps are impactful, but not why they are rated as such. The work herein will describe a new method for evaluation of relative step impact in a total synthesis route, combining a graphical representation that clearly underscores high- and low-impact steps with a distance metric based on matrix representations of intermediates capable of displaying the contributing elements involved in determining the impact of each step. This step impact evaluation metric will be applied towards several computer-generated synthetic routes towards the natural product stemoamide, in order to extract high-impact steps from several retrosyntheses for implementation into experimental routes.

The third and final section of this manuscript will detail development of a robotic platform to conduct high-throughput experiment (HTE) screens, such that valuable reactions highlighted in previous sections can be discovered in the laboratory. In an HTE screen, reactions are commonly conducted in parallel using wellplates with rectangular grids, with each well having its own unique reagents or conditions[32]. HTE is currently a rapidly developing field due to its data economy, being able to generate large amounts of organized data in a small footprint, in terms of both vessel size and material usage[33]. Many facets of chemical synthesis have benefited from adopting HTE, such as reaction condition screening[34], exploration of reaction substrate scope[35], reaction performance prediction through machine learning[36], and discovery of novel reactivities[37–39].

The standardized nature of HTE equipment readily provides opportunities for automation. Developments in robotic and computational technology has enabled systems that can conduct experiments, evaluate results, and execute the next iteration of experiments in a closed cycle[40–42]. While these systems contain immense potential to expedite advancements in various scientific fields, they currently require large capital expenditure and repurposing of laboratory space. This manuscript will focus on an automation platform utilizing the Opentrons OT-2, a compact liquid handler developed for routine bio-chemical work, but can be repurposed for use in high-throughput organic synthesis, as both disciplines share similar hardware and workflows. The OT-2's modular nature and small footprint allows its adoption by traditional chemistry laboratories without large expenditures in supporting infrastructure. The reproducibility of experiments set up in 96-wellplates by the OT-2 will be evaluated by comparison with manual setup through micropipettes and custom-made plastic scoops for poorly soluble reagents[43].

1.2 References

(1) Cayley, A. On the Analytical Forms Called Trees, with Application to the Theory of Chemical Combinations. *Rep. Br. Assoc Adv. Sci* **1875**, *45*, 257–305.

(2) Rains, E. M.; Sloane, N. J. A. On Cayley's Enumeration of Alkanes (or 4-Valent Trees). **2002**.

(3) Schiff, H. Zur Statistik Chemischer Verbindungen. *Ber Dtsch. Chem Ber* **1875**, *8*, 1542–1547.

(4) Henze, H. R.; Blair, C. M. THE NUMBER OF ISOMERIC HYDROCARBONS OF THE METHANE SERIES. *J. Am. Chem. Soc.* **1931**, *53* (8). https://doi.org/10.1021/ja01359a034.

(5) Merrifield, R. B.; Stewart, J. Morrow.; Jernberg, Nils. Instrument for Automated Synthesis of Peptides. *Anal. Chem.* **1966**, *38* (13), 1905–1914. https://doi.org/10.1021/ac50155a057.

(6) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses: Pathways for Molecular Synthesis Can Be Devised with a Computer and Equipment for Graphical Communication. *Science* **1969**, *166* (3902), 178–192. https://doi.org/10.1126/science.166.3902.178.

(7) Henze, H. R.; Blair, C. M. THE NUMBER OF STRUCTURALLY ISOMERIC ALCOHOLS OF THE METHANOL SERIES. *J. Am. Chem. Soc.* **1931**, *53* (8). https://doi.org/10.1021/ja01359a027.

(8) Perry, D. THE NUMBER OF STRUCTURAL ISOMERS OF CERTAIN HOMOLOGS OF METHANE AND METHANOL. *J. Am. Chem. Soc.* **1932**, *54* (7). https://doi.org/10.1021/ja01346a035.

(9) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3). https://doi.org/10.1021/ar500432k.

(10) 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). https://doi.org/10.1021/acs.jcim.6b00173.

(11) Hu, Q.; Peng, Z.; Sutton, S. C.; Na, J.; Kostrowicki, J.; Yang, B.; Thacher, T.; Kong, X.; Mattaparti, S.; Zhou, J. Z.; Gonzalez, J.; Ramirez-Weinhouse, M.; Kuki, A. Pfizer Global Virtual Library (PGVL): A Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information. *ACS Comb. Sci.* **2012**, *14* (11). https://doi.org/10.1021/co300096q.

(12) Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23* (11), 101681. https://doi.org/10.1016/j.isci.2020.101681.

(13) Hendrickson, J. B. Systematic Synthesis Design. IV. Numerical Codification of Construction Reactions. *J. Am. Chem. Soc.* **1975**, *97* (20). https://doi.org/10.1021/ja00853a023.

(14) Ugi, I.; Gillespie, P. Representation of Chemical Systems and Interconversions Bybe Matrices and Their Transformation Properties. *Angew. Chem. Int. Ed. Engl.* **1971**, *10* (12). https://doi.org/10.1002/anie.197109141.

(15) Ugi, I.; Stein, N.; Knauer, M.; Gruber, B.; Bley, K.; Weidinger, R. New Elements in the Representation of the Logical Structure of Chemistry by Qualitative Mathematical Models and Corresponding Data Structures. In *Computer Chemistry*; Springer Berlin Heidelberg: Berlin, Heidelberg. https://doi.org/10.1007/BFb0111463.

(16) Bauer, J. IGOR2: A PC-Program for Generating New Reactions and Molecular Structures. *Tetrahedron Comput. Methodol.* **1989**, *2* (5). https://doi.org/10.1016/0898-5529(89)90034-1.

(17) Hendrickson, J. B. Systematic Synthesis Design. III. Scope of the Problem. *J. Am. Chem. Soc.* **1975**, *97* (20), 5763–5784. https://doi.org/10.1021/ja00853a022.

(18) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2). https://doi.org/10.1039/C8SC04228D.

(19) Zhao, Q.; Savoie, B. M. Simultaneously Improving Reaction Coverage and Computational Cost in Automated Reaction Prediction Tasks. *Nat. Comput. Sci.* **2021**, *1* (7). https://doi.org/10.1038/s43588-021-00101-3.

(20) Sacha, M.; Błaż, M.; Byrski, P.; Dąbrowski-Tumański, P.; Chromiński, M.; Loska, R.; Włodarczyk-Pruszyński, P.; Jastrzębski, S. Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits. *J. Chem. Inf. Model.* **2021**, *61* (7), 3273–3284. https://doi.org/10.1021/acs.jcim.1c00537.

(21) Touré, B. B.; Hall, D. G. Natural Product Synthesis Using Multicomponent Reaction Strategies. *Chem. Rev.* **2009**, *109* (9), 4439–4486. https://doi.org/10.1021/cr800296p.

(22) Nicolaou, K. C.; Pan, S.; Shelke, Y.; Ye, Q.; Das, D.; Rigol, S. A Highly Convergent Total Synthesis of Norhalichondrin B. *J. Am. Chem. Soc.* **2021**, *143* (49), 20970–20979. https://doi.org/10.1021/jacs.1c10539.

(23) Nicolaou, K. C.; Edmonds, D. J.; Bulger, P. G. Cascade Reactions in Total Synthesis. *Angew. Chem. Int. Ed.* **2006**, *45* (43), 7134–7186. https://doi.org/10.1002/anie.200601872.

(24) Schwan, J.; Christmann, M. Enabling Strategies for Step Efficient Syntheses. *Chem. Soc. Rev.* **2018**, *47* (21), 7985–7995. https://doi.org/10.1039/C8CS00399H.

(25) Newhouse, T.; Baran, P. S.; Hoffmann, R. W. The Economies of Synthesis. *Chem Soc Rev* **2009**, *38* (11), 3010–3021. https://doi.org/10.1039/B821200G.

(26) Burns, N. Z.; Baran, P. S.; Hoffmann, R. W. Redox Economy in Organic Synthesis. *Angew. Chem. Int. Ed.* **2009**, *48* (16), 2854–2867. https://doi.org/10.1002/anie.200806086.

(27) Crossley, S. W. M.; Shenvi, R. A. A Longitudinal Study of Alkaloid Synthesis Reveals Functional Group Interconversions as Bad Actors. *Chem. Rev.* **2015**, *115* (17), 9465–9531. https://doi.org/10.1021/acs.chemrev.5b00154.

(28) Barone, R.; Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 269–272. https://doi.org/10.1021/ci000145p.

(29) Whitlock, H. W. On the Structure of Total Synthesis of Complex Natural Products. *J. Org. Chem.* **1998**, *63* (22), 7982–7989. https://doi.org/10.1021/jo9814546.

(30) Scott, K. A.; Groch, J. R.; Bao, J.; Marshall, C. M.; Allen, R. A.; Nick, S. J.; Lauta, N. R.; Williams, R. E.; Qureshi, M. H.; Delost, M. D.; Njardarson, J. T. Minimalistic Graphical Presentation Approach for Total Syntheses. *Tetrahedron* **2022**, *126*, 133062. https://doi.org/10.1016/j.tet.2022.133062.

(31) Landwehr, E. M.; Baker, M. A.; Oguma, T.; Burdge, H. E.; Kawajiri, T.; Shenvi, R. A. Concise Syntheses of GB22, GB13, and Himgaline by Cross-Coupling and Complete Reduction. *Science* **2022**, *375* (6586), 1270–1274. https://doi.org/10.1126/science.abn8343.

(32) Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.* **2017**, *8* (6). https://doi.org/10.1021/acsmedchemlett.7b00165.

(33) Wong, H.; Cernak, T. Reaction Miniaturization in Eco-Friendly Solvents. *Curr. Opin. Green Sustain. Chem.* **2018**, *11*, 91–98. https://doi.org/10.1016/j.cogsc.2018.06.001.

(34) Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.;

Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. Nanomole-Scale High-Throughput Chemistry for the Synthesis of Complex Molecules. *Science* **2015**, *347* (6217), 49–53. https://doi.org/10.1126/science.1259203.

(35) McGrath, A.; Zhang, R.; Shafiq, K.; Cernak, T. Repurposing Amine and Carboxylic Acid Building Blocks with an Automatable Esterification Reaction. *Chem. Commun.* **2023**, *59* (8), 1026–1029. https://doi.org/10.1039/D2CC05670D.

(36) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* **2018**, *360* (6385), 186–190. https://doi.org/10.1126/science.aar5169.

(37) Zhang, Z.; Cernak, T. The Formal Cross-Coupling of Amines and Carboxylic Acids to Form Sp3–Sp3 Carbon–Carbon Bonds. *Angew. Chem. Int. Ed.* **2021**, *60* (52), 27293–27298. https://doi.org/10.1002/anie.202112454.

(38) Douthwaite, J. L.; Zhao, R.; Shim, E.; Mahjour, B.; Zimmerman, P. M.; Cernak, T. Formal Cross-Coupling of Amines and Carboxylic Acids to Form Sp3–Sp2 Carbon–Carbon Bonds. *J. Am. Chem. Soc.* **2023**, *145* (20), 10930–10937. https://doi.org/10.1021/jacs.2c11563.

(39) Yayla, H. G.; Peng, F.; Mangion, I. K.; McLaughlin, M.; Campeau, L.-C.; Davies, I. W.; DiRocco, D. A.; Knowles, R. R. Discovery and Mechanistic Study of a Photocatalytic Indoline Dehydrogenation for the Synthesis of Elbasvir. *Chem. Sci.* **2016**, *7* (3), 2066–2073. https://doi.org/10.1039/C5SC03350K.

(40) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**, *583* (7815), 237–241. https://doi.org/10.1038/s41586-020-2442-2.

(41) MacLeod, B. P.; Parlane, F. G. L.; Brown, A. K.; Hein, J. E.; Berlinguette, C. P. Flexible Automation Accelerates Materials Discovery. *Nat. Mater.* **2022**, *21* (7), 722–726. https://doi.org/10.1038/s41563-021-01156-3.

(42) Jiang, T.; Bordi, S.; McMillan, A. E.; Chen, K.-Y.; Saito, F.; Nichols, P. L.; Wanner, B. M.; Bode, J. W. An Integrated Console for Capsule-Based, Automated Organic Synthesis. *Chem. Sci.* **2021**, *12* (20), 6977–6982. https://doi.org/10.1039/D1SC01048D.

(43) Cook, A.; Clément, R.; Newman, S. G. Reaction Screening in Multiwell Plates: High-Throughput Optimization of a Buchwald–Hartwig Amination. *Nat. Protoc.* **2021**, *16* (2), 1152–1169. https://doi.org/10.1038/s41596-020-00452-7.

Chapter 2 Enumeration of Amine – Carboxylic Acid Coupling Reactions Based on Matrix Encoding of Chemical Transformations.

2.1 Introduction

Amines and carboxylic acids are two widely available functional groups that are classically united through the amide coupling reaction (Fig. 2.1a, **2.1** + **2.2** → **2.3**), a tried-and-true chemistry that has become the most popular reaction for pharmaceutical explorations of chemical space[1]. To tap into this robust transformation, many research and drug discovery institutions possess large amounts of amines and carboxylic acid building blocks. We hypothesize that, by discovering new coupling chemistries between these abundant building blocks, larger amounts of chemical space can be accessed.

In our group's first study[2], the strategy was to curate a set of coupling transformations that were deemed intuitive by chemical intuition, such as coupling (**2.4**), fragmentation (**2.5**) and reduction (**2.6**). Four pairs of simple amine and carboxylic acid building blocks were computationally united at each partner's functional group atoms as well as α and β atoms, using a selection or sequence of these transformations. These virtual products spanned a wide range of physicochemical properties, suggesting that macroscopic properties such as cell permeability or metabolic stability of a product could be influenced by varying only the chemical transformation between two building blocks, in contrast to the conventional drug discovery approach of varying the building blocks themselves.

However, this curated list of transformations focused on reaction simplicity and plausibility instead of exhaustive enumeration, which excluded many conceivable and popular transformations where multiple bonds were formed or broken. These include cyclization (**2.7**), where two bonds are made between the amine–acid pair, oxidation of the bond forged during coupling (**2.8**), as well as less intuitive instances where at least one of the substrates is fragmented, such as addition (**2.9**), rearrangement (**2.10**), insertion (**2.11**) and metathesis (**2.12**). In this subsequent study, we aim to perform a much more extensive enumeration process that will include all reactions presented above, as well as any other transformations that are theoretically possible without violating the octet rule.

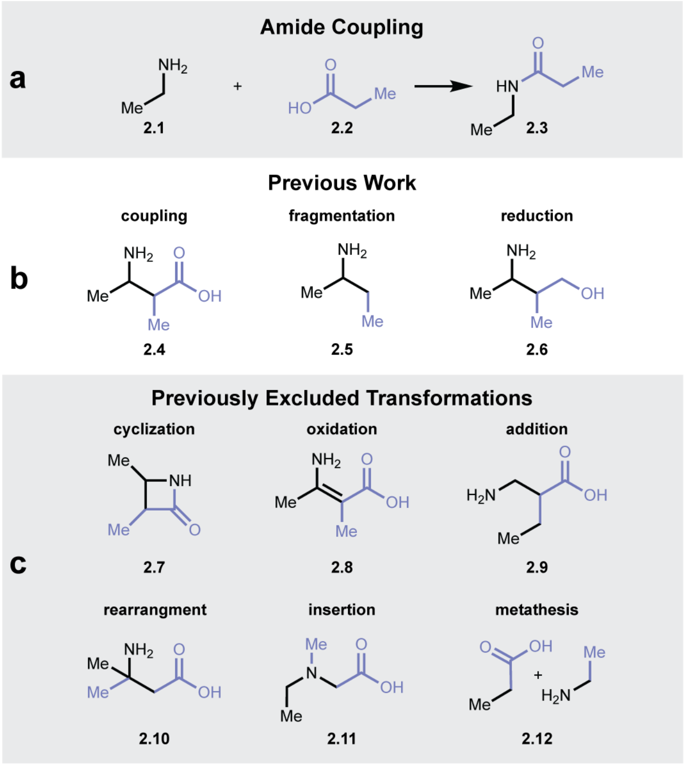


Figure 2.1. Diverse amine–carboxylic acid transformation products. **a.** Given an amine **2.1** and carboxylic acid **2.2**, the most popular transformation that unites this pair of building blocks produces the amide **2.3**. **b.** Coupling products arising from a curated subset of chemical transformations charted by our prior work. **c.** Examples of transformations not appearing in our previous publication, that we wish to consider in this manuscript.

2.2 Background of reaction encoding via matrices

The application of graph theory concepts towards organic chemistry, wherein molecules are represented as molecular graphs, with atoms as nodes and bonds as edges, have been performed as early as 1875 to enumerate selected molecule classes, such as branched alkanes[3,4], alcohols[5,6], and cyclic carbon skeletons[7]. These methods later evolved to matrix encoding of molecular graphs[8] as well as molecular reactions, most notably by Ugi, Dugundji and coworkers[9–13], who established the concept of a *be*-, or bond-electron matrix to encode molecular structure by denoting the distributions of bonds and electrons, and the difference between two *be*-matrices to be the reaction matrix, encoding information of bond and electron flow. These tools were utilized for reaction enumeration in an orthogonal direction to our work, selecting one particular reaction matrix while permuting identities of atoms at which the transformations occur[14].

Other variants of matrix encoding have been reported, such as bond/edge matrices[15,16], generalized graph matrices[17,18], 3D molecular graphs[19–21], atom-bond connectivity matrices[22–26], and other graph theoretical matrices[27,28]. In contemporary literature, matrix encoding of molecules and chemical transformations have been applied towards prediction of molecular properties and reaction outcomes[29,30], and enumeration of chemical spaces bound by number of bond transformations[31], ring count[32], or atom count[33].

2.3 Mathematical basis of coupling reaction enumeration

In our exploration of the amine–carboxylic acid coupling space, we accessed that the upper bound to the occupants of this space is the total number of ways the building blocks' constituent atoms, or a subset thereof, can be covalently attached without violating valency rules. We chose to only consider structures where all atoms do not have formal charges, although species with charged atoms, such as nitro groups or quaternary ammonium ions, are conceivable.

Figure 2.2a demonstrates a simple example with the coupling of ethylamine (**2.1**) and propanoic acid (**2.2**) to form amide **2.3**. First, a matrix is generated for the starting system. During this reaction, the bond order between N6 and C5 increases by one, while the bond order between C5 and O8 decreases by one. Addition of this transformation matrix to that of the starting materials then gives the amide product **2.3**. This same transformation matrix can be obtained by first separating this two-carbon amine and three-carbon carboxylic acid system into constituent atoms (Fig. 2.2b), obtaining amide **2.3** by adding entries into a blank adjacency matrix, and then subtracting the adjacency matrix of the reactants from the adjacency matrix of the products (Fig. 2.2c), in analogy to the work of Ugi[9], Schneider[34] and other colleagues.

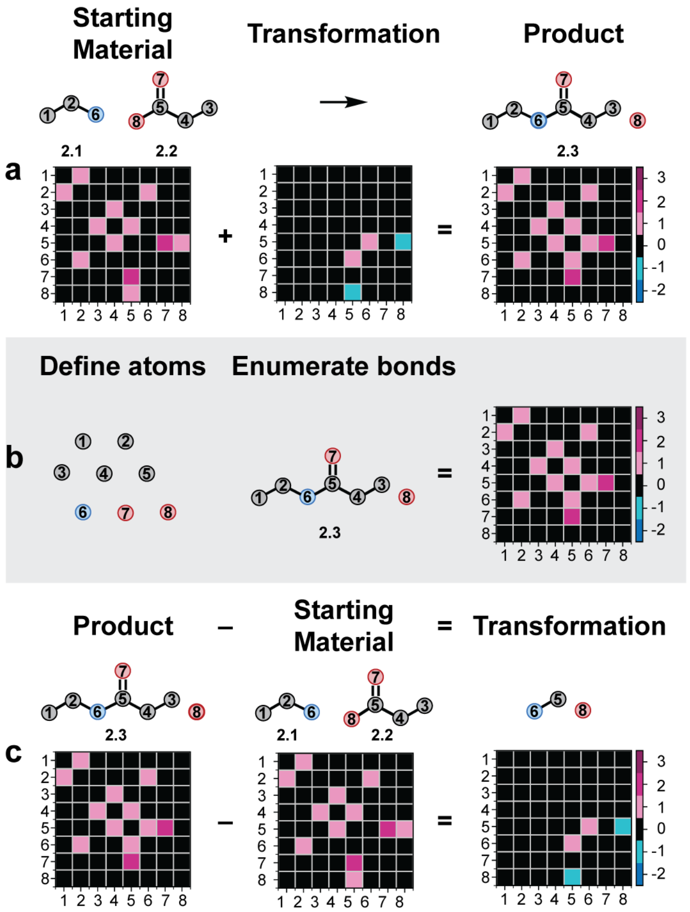


Figure 2.2. Matrix enumeration in coupling product generation. **a.** Demonstration of matrix-encoding of reaction from Figure 1a. In the molecular diagrams, white circles represent carbon, blue circles nitrogen, and red circles oxygen. The numbers in the adjacency matrix correspond to the atom indices in the cartoon atoms, while the color represents bond order. **b.** Workflow for exhaustive generation of amine–acid coupling products. **c.** The difference between the product matrix and starting material matrix is the transformation matrix.

The procedure to exhaustively generate all amine – carboxylic acid coupling products and transformation matrices is as follows:

1. All heavy atoms are first assigned indices. The exact assignment of atoms to indices does not affect the result, but matrix visualizations will appear more intuitive if adjoining atoms in the molecule are given indices that put them next to, or close to each other in the matrix.
2. Compute the matrix representation of this initial system, for generation of transformation matrices further down the workflow. Each bond between heavy atoms is encoded as an integer, in the position corresponding to its neighboring atoms' indices. For example, in Fig. 2.2a, since atoms #2 and #6 have a single bond between them, then the matrix entries (2,6) and (6,2) are set to 1. Similarly, the (5,7) and (7,5) entries are set to 2. In our work,

four starting material representations are generated, for each permutation of hybridization of the α and β carbons (Figure 2.3).

3. Determine bonding parameters for all atoms. Every atom has two parameters: the maximum total bond order, *t,* originating from it, and the highest order for an individual bond, *b*. For example, a neutral nitrogen atom has (*t, b*) = (3,3), while a neutral carbon has (*t, b*) = (4,3) since, while carbon can have a total bond order of 4, it is not permitted to make quadruple bonds.
4. Determine the identity of the first row's corresponding atom. Generate all possible rows that sum to *t* or below (since bonds to hydrogen are implied), with each entry having a maximum value of *b*, and the first entry being 0 (since an atom cannot bond with itself).
5. For each generated matrix, the first entry of the second row is set to the second entry of the first row, and the second entry is set to 0.
6. Using the second row atom's values of *t* and *b*, generate all possible remaining variants of the second matrix row.
7. Repeat steps 5 and 6 to generate subsequent rows, where the *n*th row is initialized by vertically stacking the previously generated rows into a rectangular matrix, copying the *n*th column into the *n*th row, then generating all permitted permutations of the *n+1*th element and beyond which the *t* and *b* values permit.
8. Following this algorithm, the last row of each matrix is a copy of the penultimate column with an appended 0.
9. Obtain the set of transformation matrices from each starting material pair, via subtracting a matrix generated in step 2 from each product matrix resulting from steps 4-8.

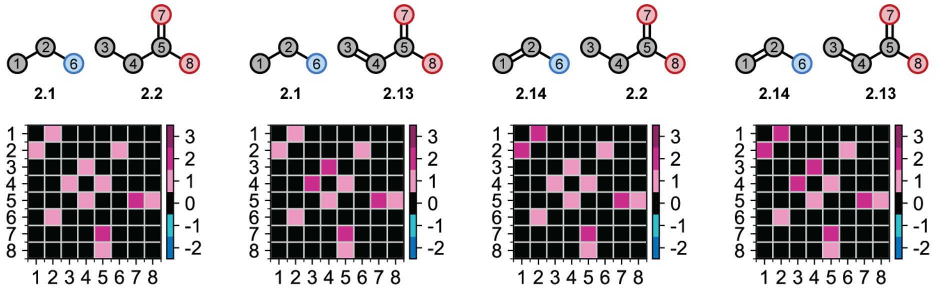


Figure 2.3. The four starting material hybridization permutations, and their matrix representations. From left to right: *sp*^{3} amine **2.1** and *sp*^{3} acid **2.2**, *sp*^{3} amine **2.1** and *sp*^{2} acid **2.13**, *sp*^{2} amine **2.14** and *sp*^{3} acid **2.2**, and *sp*^{2} amine **2.14** and *sp*^{2} acid **2.13**.

2.4 Results of amine–carboxylic acid coupling reaction enumeration

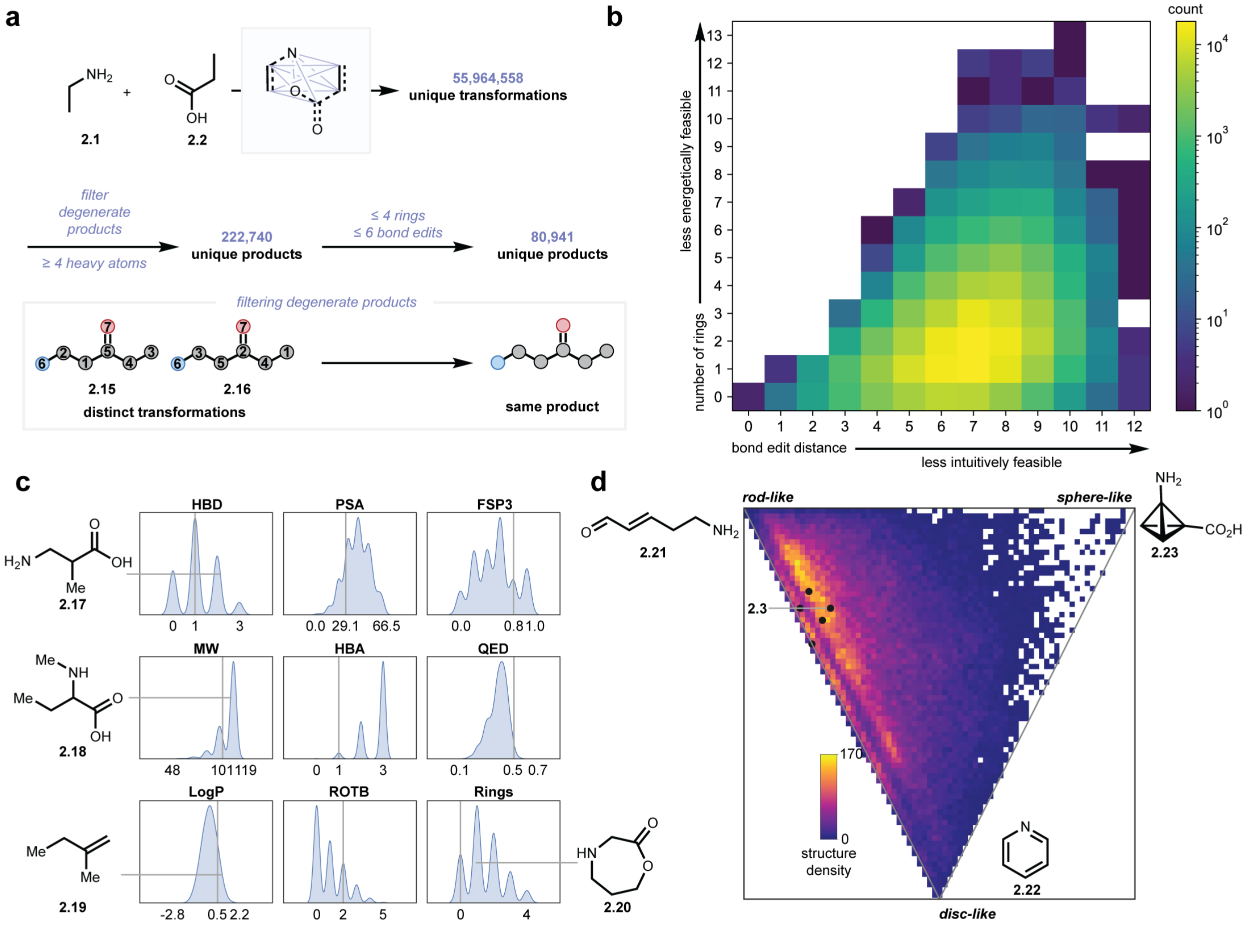


Figure 2.4. Results and analysis of amine–carboxylic acid reaction enumeration. **a.** Schematic of enumeration from amine **2.1** and acid **2.2** to yield 56 million unique transformation matrices, which are filtered first to 222,740 unique products assuming carbon and oxygen atoms are degenerate, and further to 80,941 unique products after eliminating highly improbable structures. **b.** Two-dimensional histogram showing distribution of ring count and bond edit distance of the initial 222,740 products. **c.** Kernel density estimate (KDE) plots of various physiochemical properties of the filtered amine-acid coupling system with 80,941 structures, along with selected products. The respective property of the classic amide is shown by the vertical grey line. HBD = hydrogen bond donor, PSA = polar surface area, FSP3 = fraction sp^3-atoms, MW = molecular weight, HBA = hydrogen bond acceptors, QED = quantitative estimate of drug-likedness, LogP = partition coefficient, ROTB = number of rotatable bonds, Rings = number of rings. **d.** Principal Moment of Inertia (PMI) ratio distributions of all products from the expanded enumeration.

2.4.1 Overview of product count and initial filtering for energetic plausibility

The structure enumeration algorithm generated an initial count of 55,964,558 transformation matrices (Fig. 2.4a)[35], considering that only eight atoms were being united. Since the two oxygen atoms on a carboxylic acid are almost always chemically equivalent, the total number of products can be reduced by approximately half to 23,829,176, unless isotopic labelling of ^{16}O versus ^{18}O is incorporated. Setting all 5 carbon atoms as degenerate, and only considering products with four or more heavy atoms, further simplifies this space to 222,740. Though a relatively drastic 100-fold reduction, this quantity of coupling products remains remarkable, since only a maximum of eight atoms were incorporated in each structure.

Within this vast structural space, many products contained motifs that are energetically improbable, or structurally distant from the simple amine–acid building blocks. To limit the inclusion of such improbable structures, we eliminated any structure with more than 4 rings, or requiring more than 6 bond edits from the amine–acid substrates (Fig. 2.4b). This brought the structure count to a final 80,941 products.

2.4.2 Distribution of physicochemical properties

Molecular properties of the enumerated set of 80,941 amine–acid products were computed with the RDKit package[36], and their distributions visualized as kernel density estimate (KDE) plots (Fig. 2.4c). For comparison with the amide coupling, properties of the corresponding sp^3–sp^3 amide (**2.3**) are plotted as a vertical line in each panel. Expansion of the reaction space has a variety of effects on molecular properties compared to our earlier map[2]. While the amide product **2.3** has one hydrogen bond donor (HBD), the matrix enumeration set was allowed to break the carboxylic acid moiety into two alcohol groups, each being an HBD, accounting for the small peak at HBD = 3. Molecular weight (MW) and number of hydrogen bond acceptors (HBA) skewed larger when the reaction space was expanded. Both observations can be attributed to the increase in structures that incorporate all carbon, nitrogen, and oxygen atoms (**2.17, 2.18, 2.20**). Since the number of possible structures increases with the number of atoms involved, expanding the reaction set affords an increase in the number of relatively massive structures, accounting for the shift in distribution to larger MW. The distribution of MW also has a long tail towards structures with low mass, as fragmentation transformations result in products with fewer atoms than the starting materials. The same reasoning holds for the large peak at HBA = 3, since more product substructures can incorporate all three polar atoms, as in **2.17**. Other properties distribute more to lower values, such as logP, due to polar small molecules, and number of rotatable bonds (ROTB), due to highly rigid caged structures.

2.4.3 Distribution of molecular shape

To examine the shape diversity of the 80,941 matrix-enumerated products, a principal moment of inertia (PMI) ratio analysis[37] was performed (Fig. 2.4d). The enumerated reaction space covers a high diversity of three-dimensional shapes, including rod-like molecules such as **2.21**, disc-like molecules such as **2.22**, and sphere-like molecules such as **2.23**. The *sp*^{3}–*sp*^{3} amide coupling product **2.3**, being a nearly linear molecule, sits near the upper left corner of the PMI plot. Upon application of matrix enumeration, the available molecular shapes cover a much larger portion of the PMI plot, suggesting that a wider diversity of molecular shapes can potentially be achieved by expanding the library of amine–acid coupling reactions.

2.5 Application towards examination of retrosynthetic opportunities.

Our dataset of virtual amine–acid coupling products can also be used as a retrosynthetic strategy to disconnect complex molecules for total synthesis. Through substructure searches of our enumerated products within a desired synthetic target, we can inform potential retrosynthetic disconnections for synthetic planning of complex molecules, as outlined in Figure 2.5. Lactone **2.24**, a structure generated via transformation enumeration, can be formed by cyclization, where the acid oxygen and β carbon of propenoic acid (**2.13**) form σ bonds to the amine β carbon of ethylamine (**2.1)** (Fig. 2.5a). Therefore, in any target molecule that contains **2.24** as a substructure, such as noscapine (**2.25**), the transformation in Figure 2.5a can present a retrosynthetic disconnection, simplifying the target into smaller building blocks (Fig. 2.5b).

Within each drug molecule, the result of searching through all 80,941 amine–acid coupling structures can be visualized as a chord diagram (Fig. 2.5c and 2.5d). Target molecules to be disconnected lie on the black band, and matrix enumeration products on the checkered band. Each chord represents one disconnection as presented in Fig. 2.5b. The high degree of connectivity demonstrates the many opportunities for retrosynthetic simplification that arise through the invention of new reactions. For example, in addition to the **2.26**–**2.27** pair, noscapine can undergo another disconnection pathway into *sp*^{3} acid **2.28** and amine **2.29**. In some products, a single substructure can present more than one disconnection mode, such as athamontanolide (**2.30**) disconnecting into acrylic acid (**2.31**), and either amine **2.32** or **2.33**. We propose that this disconnection search method can complement current retrosynthetic algorithms, by screening many disconnection modes for readily available or easily synthesizable building blocks, then presenting the coupling reaction to be developed in an experimental setting.

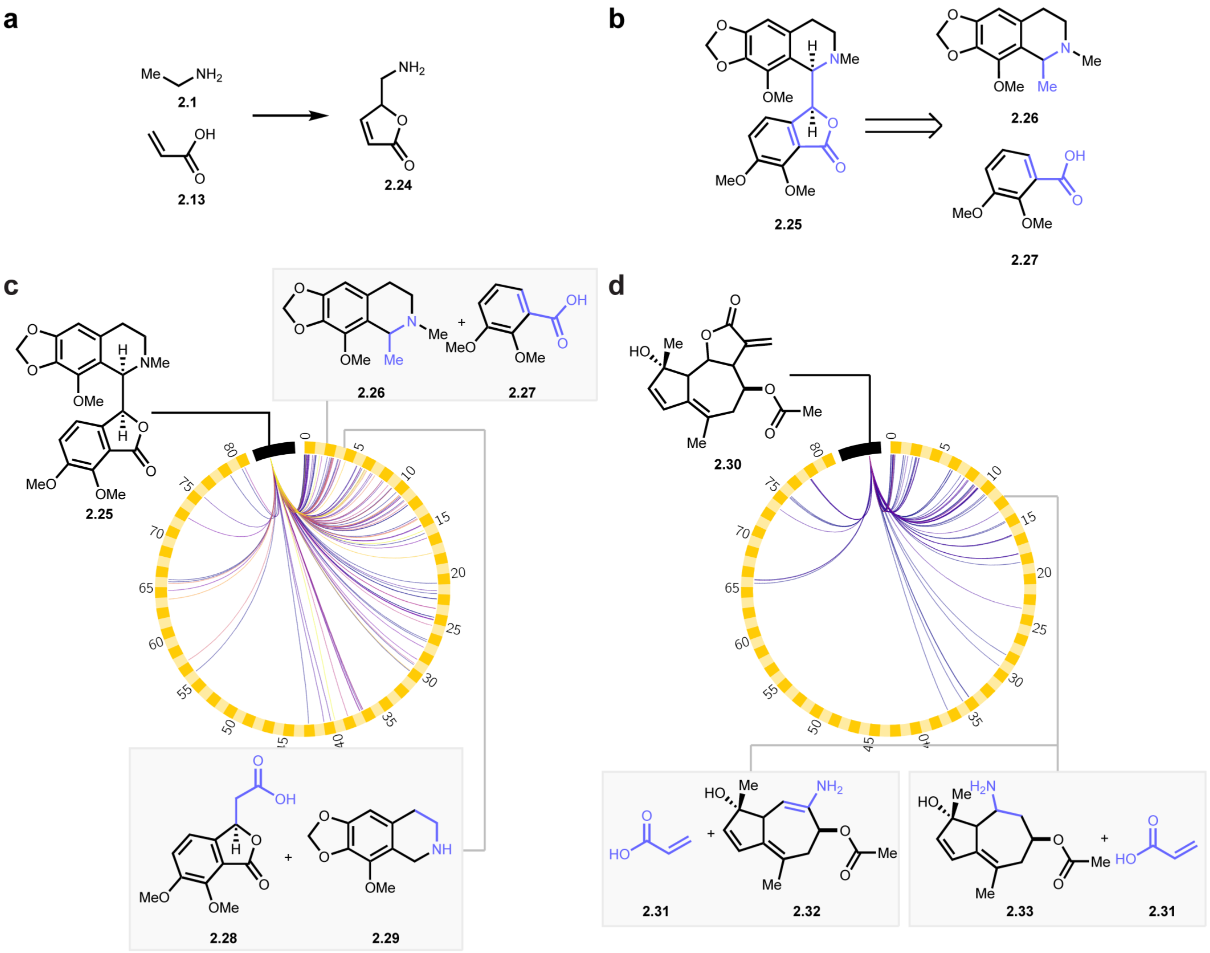


Figure 2.5. Analysis of retrosynthetic opportunities using products generated through virtual amine–acid coupling. **a.** Amine **2.1** and acid **2.13** couple to give lactone **2.24**. **b.** By reversing the transformation in 5a, noscapine (**2.25**) is disconnected into amine **2.26** and acid **2.27**. **c**. Chord diagram visualizing retrosynthetic disconnection of noscapine (**2.25**). **d**. Chord diagram visualizing retrosynthetic disconnection of athamontanolide (**2.30**). A list of all substructures found in the two drugs can be found in Appendix A.

A search of all 222,740 enumerated amine–acid coupling products was conducted within the Drugbank[38] database. To effectively visualize this result within a high dimensional chemical space, a two-dimensional Uniform Manifold Approximation and Projection (UMAP)[39] was applied to 2,048-bit Morgan fingerprints[40] of all structures, and the resultant embedding visualized in Fig. 2.6. Each dot represents one structure, and the color represents its frequency of occurrence within Drugbank. Most structures commonly occurring within drugs were observed to gather largely in two neighboring clusters (yellow dots in the vicinity of **2.34**), demonstrating the vastness of structural space that remains unexplored by pharmaceuticals and could be made accessible through novel amine–acid couplings. Notable "drug-like" transformations from other clusters include those that generate rings such as pyridine (**2.22**) and furan (**2.36**), suggesting that an amine–acid pyridine synthesis would be valuable in drug discovery. Amine–carbonyl condensations are a popular approach towards pyridines[41,42], including an amino acid fragmentation-reconstruction reaction, hence it is conceivable that several amine–acid pyridine syntheses could be developed to quickly access an even wider range of substituted pyridines.

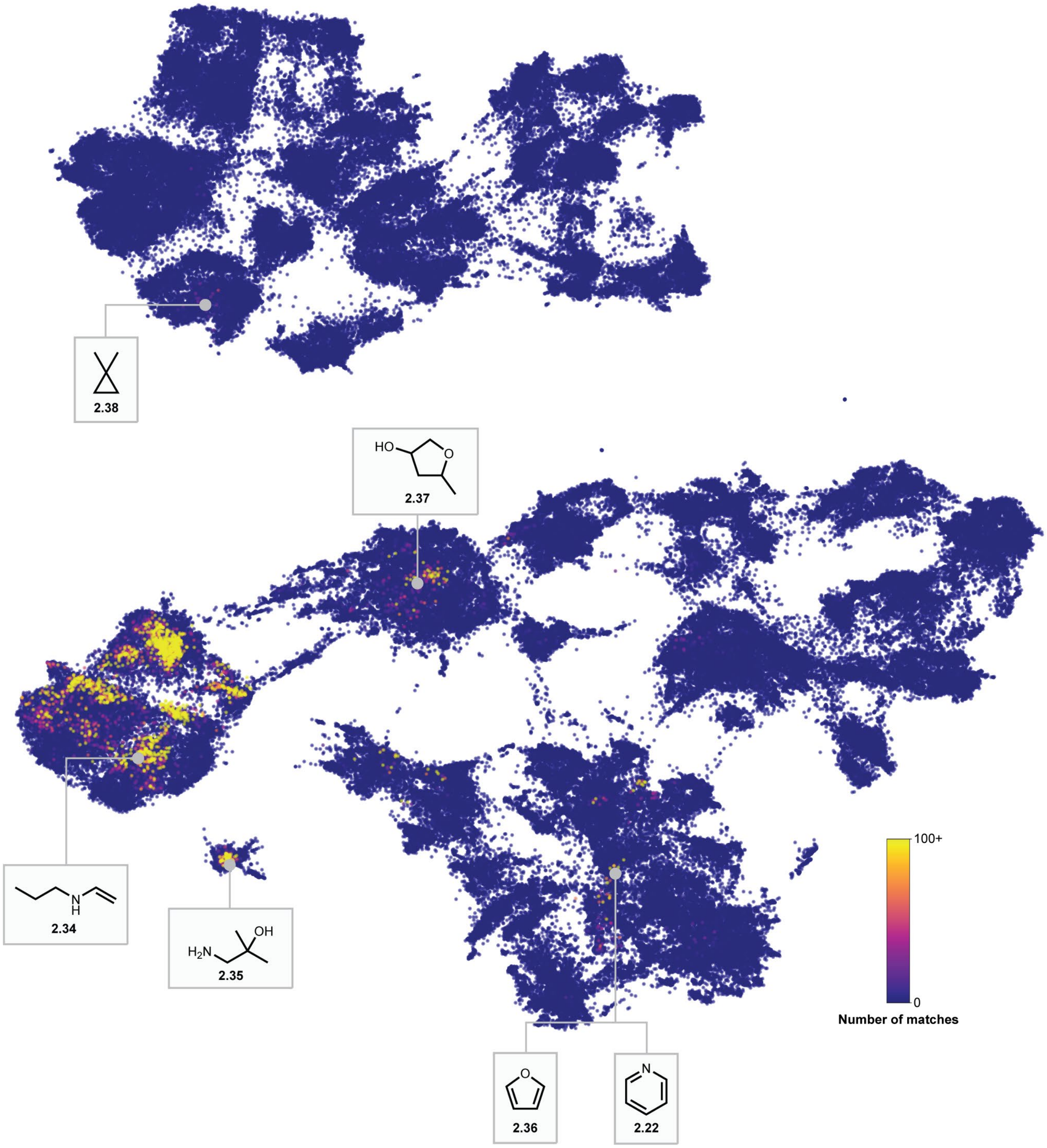


Figure 2.6. UMAP projection of Morgan fingerprints computed from 222,740 enumerated products. Dots are colored by the number of product substructure matches in the Drugbank database, with majority of product substructures appearing as unexplored chemical space. Sample structures from clusters containing products who appear as substructures in Drugbank (**2.22**, **2.34-2.38**) are displayed. A list of the top 100 most frequent substructures found in the Drugbank is located in Appendix A.

To highlight impactful amine–acid reactions for development, the most frequently occurring matrix-enumerated coupling product substructures in DrugBank are displayed in Fig. 2.7a. The most abundant substructure containing only carbon is the C–C–C–C motif **2.39**. Meanwhile, the sp^3–sp^3 C–N coupling **2.42** and the sp^3–sp^3 C–O coupling motif **2.45** and are the most abundant in structures containing only carbon and nitrogen, and carbon and oxygen respectively. Together, these findings suggest that these three transformations could be valuable in drug discovery and natural product synthesis.

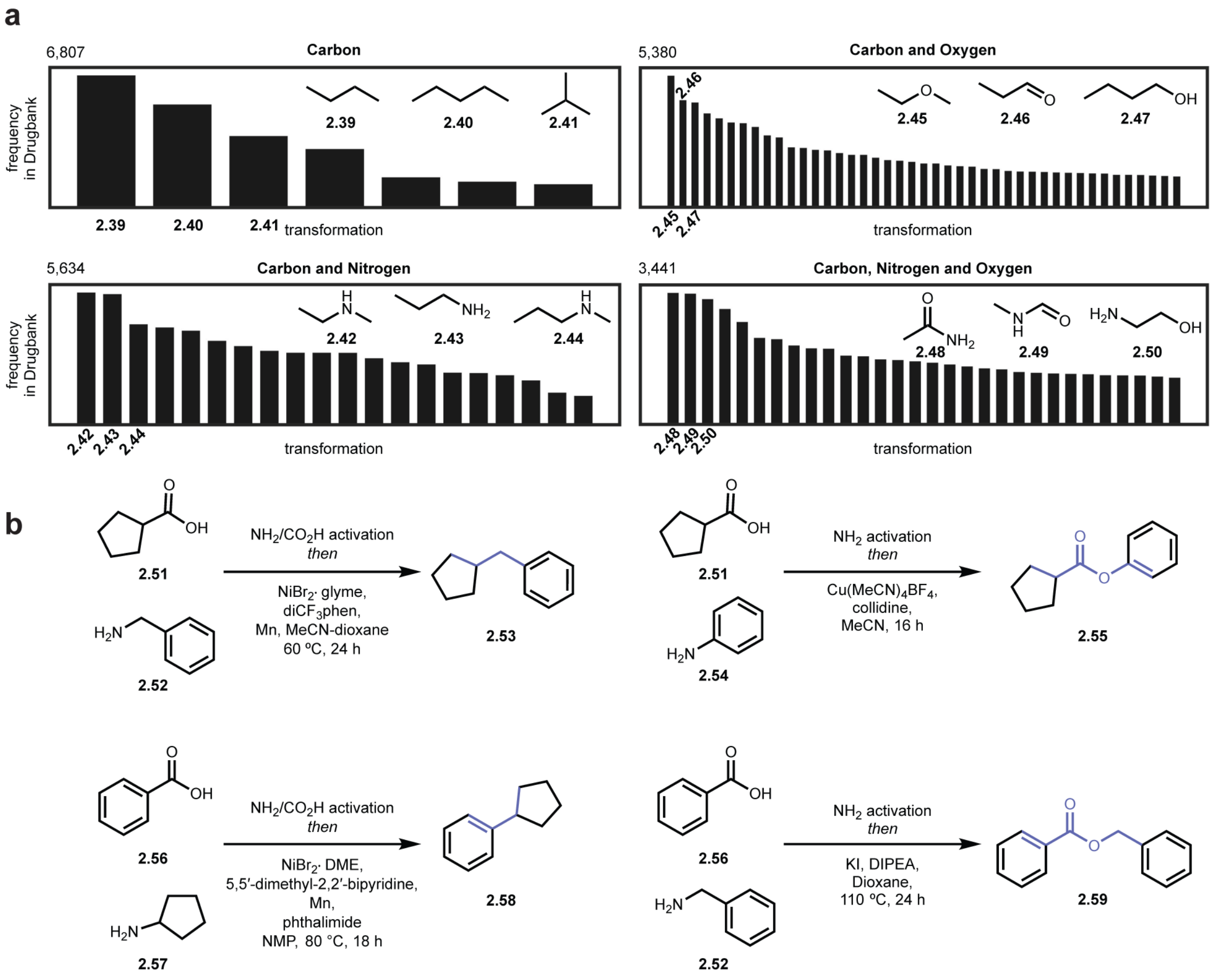


Figure 2.7. High-impact reactions recommended for discovery. **a**. Frequencies of the top 100 most abundant substructure matches in DrugBank, plotted as histograms categorized by their elemental makeup. The top three most abundant structures for each group are shown (**2.39**-**2.50**). **b.** Experimental conditions for four amine–carboxylic acid couplings subsequently discovered by our group. sp^3 acids (**2.51**) can couple with either sp^3 amines (**2.52**) to form a C–C bond through deamination and decarboxylation (**2.53**), or with sp^2 amines (**2.54**) through deaminative esterification (**1.55**). The latter conditions are also capable of coupling sp^2 acids. sp^2 acids (**2.56**) can couple with either sp^3 amines (**2.57**) to form a C–C bond through deamination and decarboxylation (**2.58**), or undergo deaminative esterification (**2.59**), which is also robust towards sp^3 acids.

The most abundant substructures with C, N and O atoms contain fragments of the amide bond (**2.48**, **2.49**), which once more underscores the prevalence of the amide coupling in drug discovery, but also opportunities in accessing complementary chemical space with amine–acid couplings that do not produce the amide, yet preserve C, N and O atoms. Following this guidance, our group has experimentally realized several amine–carboxylic acid couplings (Figure 2.7b), including *sp^3* amine–*sp^3* acid C–C coupling[43], *sp^2* amine–acid esterification[44], *sp^3* amine–*sp^2* acid C–C coupling[45], and *sp^3* amine–acid esterification[46].

2.6 Application towards virtual late-stage diversification

So far, our analysis of physicochemical property modulation has been restricted to that of small amine–acid coupling pairs. To explore the extent of this effect on larger building blocks, we applied the matrix-derived amine–acid transformation enumeration method towards the virtual late-stage diversification of druglike molecules. Traditionally, this has been achieved through the selection of alternative building blocks[47–50],which are then united with the substrate through popular reactions such as the amide coupling, Suzuki coupling and Buchwald-Hartwig coupling. To investigate how we can complement this traditional strategy, we applied matrix-derived amine–acid transformations to two druglike molecules, following the workflow outlined in Fig. 2.8a:

1. All transformation matrices are applied to the druglike molecule (**2.60**) and building block (**2.61**) by editing bond orders of the atoms at the reaction center, as instructed by the transformation matrix.
2. Since there are substitutions at some α and β carbons, some transformations will lead to products that disobey the octet rule. These structures are screened and removed through analysis with RDKit.
3. Products with high ring strain are removed from the data set.
4. Other structural elements are checked for, and preserved or excluded depending on the identity of the coupling molecule pair.
5. Matrices that result in valid coupling products are extracted.
6. The set of matrices in step 5 are applied to a simple amine–acid pair (**2.1** and **2.2**) so that they undergo the same transformation as the druglike coupling pair (cf. step 1).

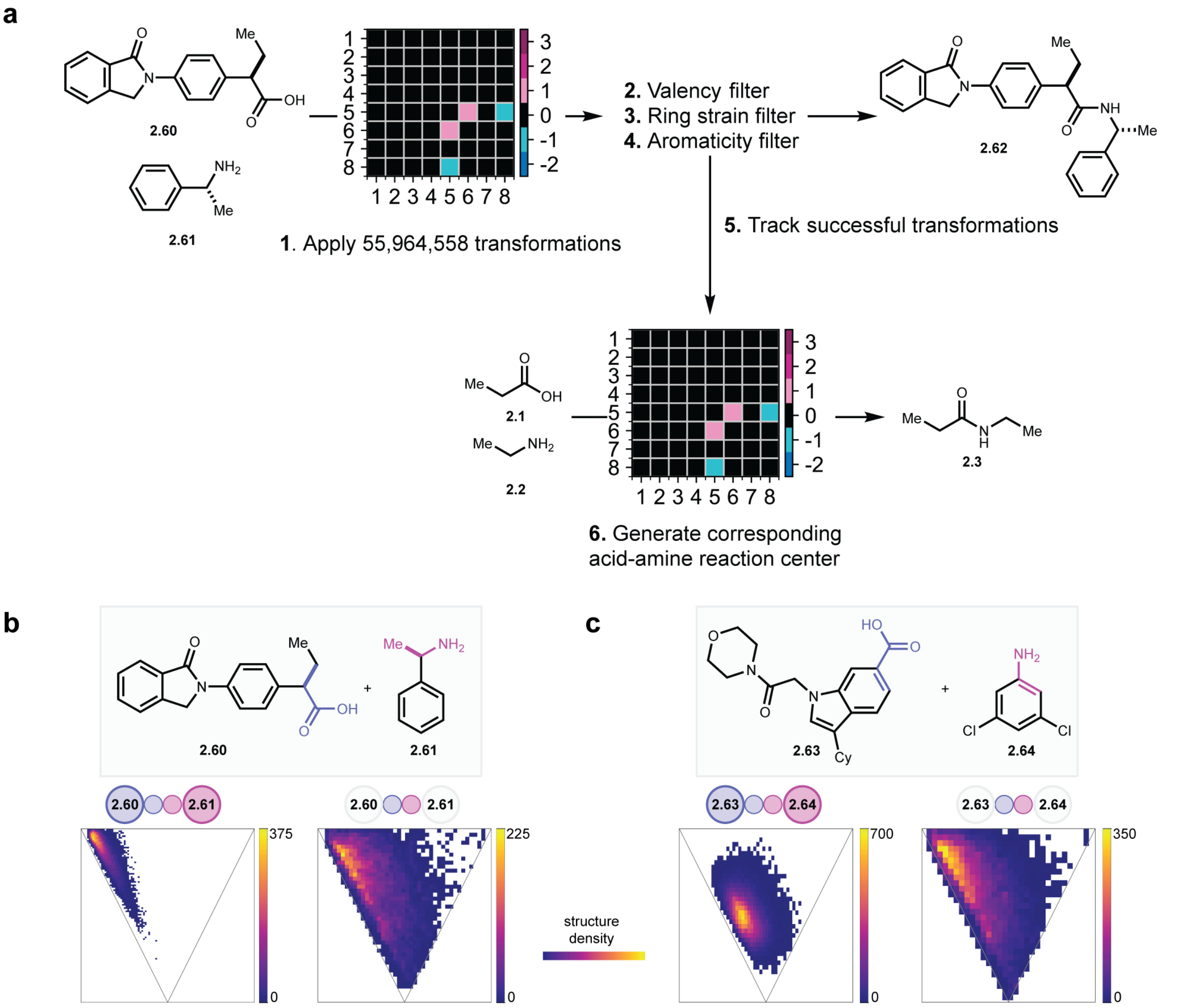

a
2.60
2.61
1. Apply 55,964,558 transformations
2. Valency filter
3. Ring strain filter
4. Aromaticity filter
2.62
5. Track successful transformations
2.1
2.2
6. Generate corresponding acid-amine reaction center
2.3
b
c
2.63
2.64
structure density
375
225
700
350
0

Figure 2.8. Process and results of virtual late-stage diversification study. **a.** Workflow of virtual late-stage diversification and generation of corresponding reaction centers. **b.** PMI ratio plots showing shape space distribution of reaction enumerated late-stage diversification of two drug-like molecules, one showing an sp^3–sp^3 coupling between **2.60** and building block **2.61**, and **c.** another showing an sp^2–sp^2 coupling between **2.63** and building block **2.64**. For each pair of plots, the left shows the distribution of full coupling products, while the right shows the distribution of only the atoms at the reaction center (*cf.* Fig. 2.4d).

This workflow was applied to two druglike molecules – one containing an sp^3-acid (**2.60**) (Fig. 2.8b) and one containing an sp^2-acid (**2.63**) (Fig. 2.8c), with a simple amine of the corresponding hybridization (**2.61** and **2.64**). Steps 3 and 4 were implemented with the goal of aligning our results with desired outcomes of late-stage diversification in the laboratory. Products that lost a significant substituent in **2.60** and **2.61** removed, while the **2.63–2.64** coupling set only retained products that retained a six-membered ring at the reacting atoms. Steps 5 and 6 are targeted towards investigating the effect of molecule size on property modulation brought about by varying transformation mode.

For each transformation product set between the druglike building blocks, represented by the larger circles in Figs. 2.8b and 2.8c, a complementary set of only the reaction center is also produced, represented by the smaller colored circles in Fig. 2.8a and 2.8b. The substrate has a considerable effect over the modulation of molecular shape among the products, as illustrated in Fig. 2.8b and 2.8c. Though the reacting atoms of the carboxylic acid and amine functional groups span a more diverse shape space, the starting materials are larger (for example, **2.60**/**2.61** versus **2.1**/**2.2**) so they exert more influence on the final product's shape. Between **2.60** and **2.61**, the aggregate PMI ratios skew towards the 1D-2D line but still distribute towards the 3D region of the PMI due to the substituent's higher linearity compared to **2.63** and **2.64**. The products in the latter set were also filtered to those retaining the aromaticity of the benzoic acid and aniline moieties, so there are fewer molecules and less coverage in this sp^2–sp^2 pairing.

The influence of substrate on physicochemical property modulation was also analyzed for the **2.60-2.61** pair, and visualized as 2D density plots in Fig. 2.9. The modulation of each property with regard to the amide (**2.62**) is plotted on the x-axis for the whole molecule, and y-axis for the reaction center. Molecules that lie on the lower left-upper right diagonal exhibit the same modulation for that property, regardless of whether the whole molecule, or only the reacting atoms are considered. Among the properties computed, HBA and HBD vary in similar amounts, since each datapoint represents the same transformation, as HBA and HBD are context-independent and additive properties. The parallel diagonal bands in MW, and some extent LogP reflect a divergence between transformations that lose atoms carrying functional groups (the large functional groups on the α-carbons of **2.60** and **2.61**), while movement along the bands reflect simultaneous transformations that occur on the other atoms (such as the acid oxygens). Meanwhile, ROTB exhibits a broader spread, as it is influenced by structural motifs on both the reacting atoms and the other atoms of the building blocks. The change in fraction sp^3 atoms (FSP3) varies more discretely for the reaction center, since there are only 8 atoms in the model system, hence the fraction can only change in multiples of 0.125. Lastly, quantitative estimate of drug-likeness (QED), being an aggregate property, shows no clear trend, as evidenced by differences in the distribution of each property whether the full product or only the reacting atoms are considered.

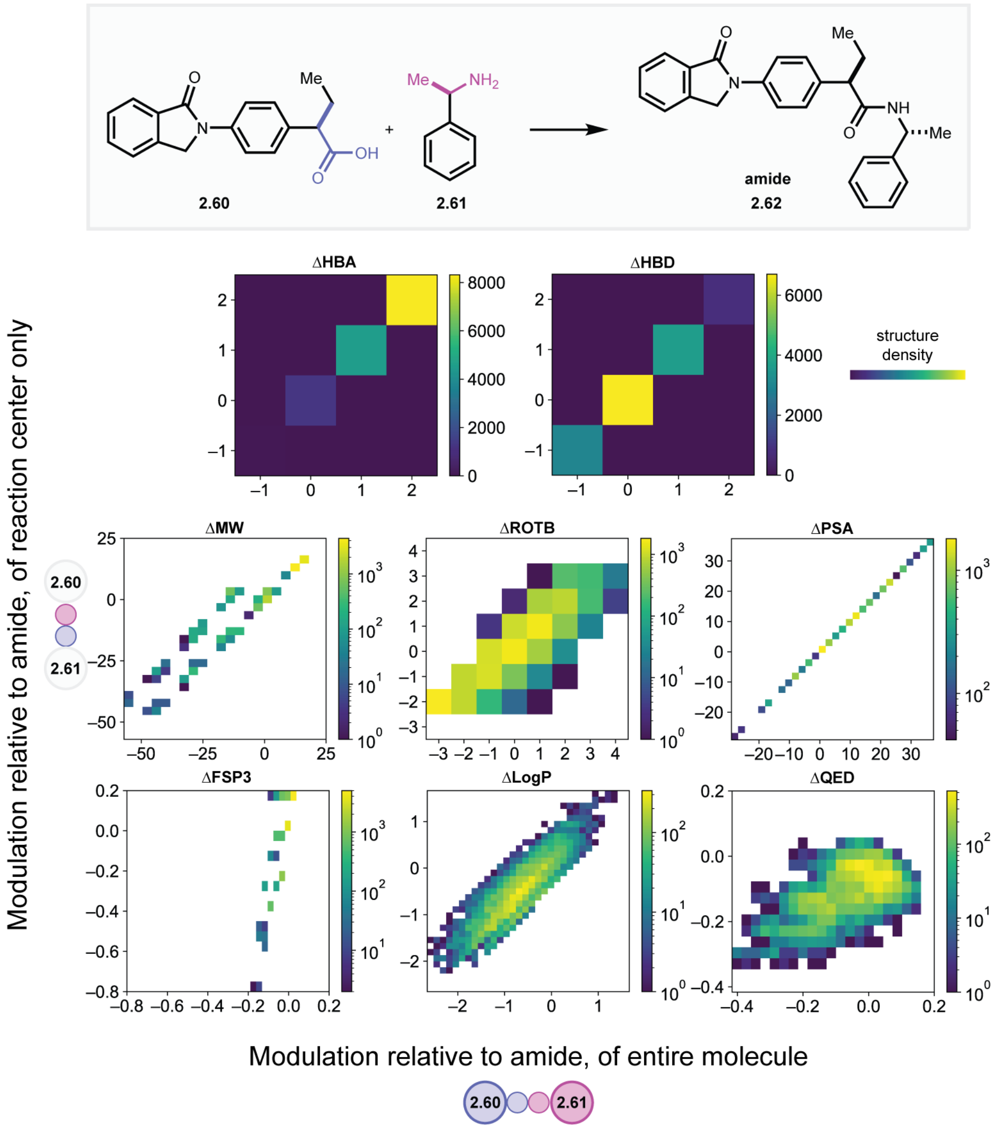


Figure 2.9. 2D density plots of physicochemical property modulation within the **2.57**–**2.58** amine–acid system. The x-axes show modulation of the entire molecule relative to the amide (**2.59**), while the y-axes show modulation of only the atoms at the reaction center.

2.7 Virtual docking of compound libraries generated through diverse amine–acid coupling reactions

Finally, we sought to evaluate the performance of matrix-derived late-stage diversification in practice. One amine–acid pair (**2.65** and **2.66**) resembling known inhibitors of the SARS-CoV-2 main protease (M^{pro}) was designed (Figure 2.5a). Two orthogonal virtual libraries were created, one from enumerating amine–acid transformations between the pair via the workflow presented in section 1.5, and the other from performing only amide coupling of **2.62** with diverse primary and secondary amines retrieved from the PubChem database[51], filtered to less than 13 heavy atoms, no elements other than H, B, C, N, O, F, P, S, Cl, Br and I, and less than two total nitrogen and oxygen atoms, to maintain similar polarity between the two libraries. A UMAP analysis performed on 2048-bit Morgan fingerprints of sampled structures in both libraries showed complementary coverage of chemical space with little overlap, illustrating that invention of diverse amine–acid reactions will allow access of chemical space previously untapped by modifying building blocks alone. Docking of both virtual libraries was performed across three M^{pro} conformers (Fig. 2.10c). Across all conformers, the library generated via diverse reactions consistently achieved docking energies with a narrower spread than the library from diverse amines, while also performing more favorably.

This study was repeated with two additional molecules – finasteride (Figure 2.10d, **2.67**) docking with Steroid 5-alpha-reductase (7BW1)[52], and diethyltoluamide (DEET) (Figure 2.10e, **2.68**) with Odorant Binding Protein 1 (3N7H)[53]. Both small molecules were disconnected at their amide bond, and a virtual library generated through recombining the resultant amine–acid pair with diverse transformations. As both acid building blocks possess ring systems, the transformation products were filtered to structures that preserved the ring system for the finasteride data set, and aromaticity of DEET.

Compared to the M^{pro} inhibitor study, the library generated by combining one building block pair with diverse transformations did not produce a significantly different docking score distribution, but maintained a narrower spread. Interestingly, both methods of virtual library generation produced better average scores than the original amides, as indicated by the vertical lines in Figs. 2.10d and 2.10e. These studies reinforce our hypothesis that conducting diverse reactions between two building blocks allows for finer tuning of physicochemical properties than the traditional method of varying the building block itself. A future approach to *in silico* screening may be to first determine the target compound's constituent building blocks, followed by probing the pair's coupling space to determine the most favorable reaction to unite them with.

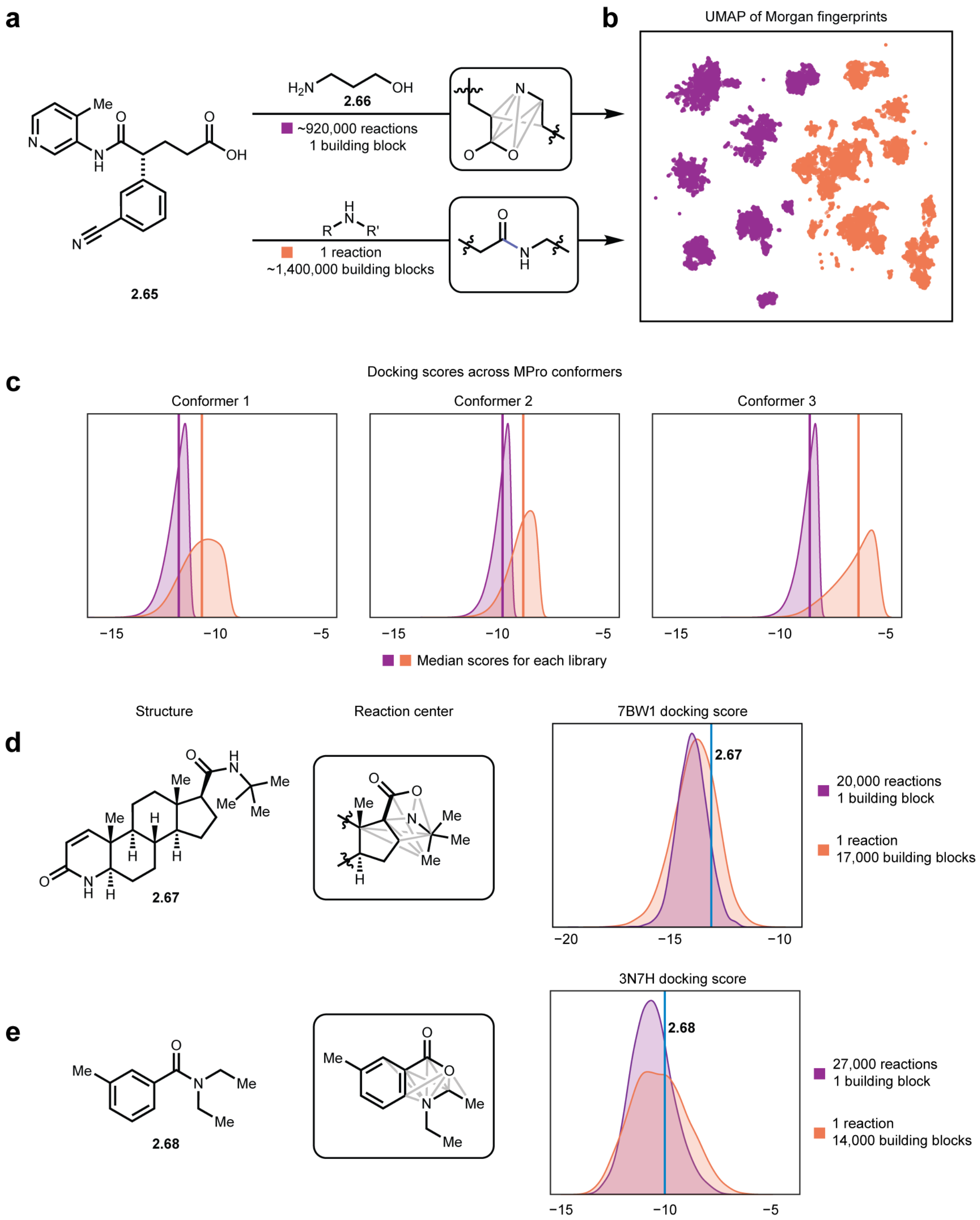


Figure 2.10. Workflow and results of docking studies. **a.** Workflow in generating two orthogonal virtual libraries. **b.** Two-dimensional UMAP projection of the combined chemical space, using 1,024-bit Morgan fingerprints, colored by the library in which each molecule belongs in. **c.** Docking energies of both libraries across three M^{pro} conformers. **d.** Same workflow as 5a applied to finasteride (**2.67**), a schematic representation at the atoms at which transformations were permitted to occur (middle), and distribution of docking scores (right). Filters were applied to include only products that preserved the local ring structure. **e.** Same workflow as 5a applied to DEET (**2.68**).

2.8 Conclusions and Future Work.

We have herein developed a method to computationally enumerate a vast array of transformations between a simple amine–carboxylic acid building block pair, producing a broad chemical space for exploration. Due to the abundance of amine and carboxylic acid building blocks, each amine–acid coupling identified and discovered represents an additional method to couple these moieties, accessing novel properties and structures.

The database of product structures, as well as their transformation recipes, can be applied towards simple retrosynthetic searches, or producing coupling libraries between larger molecules that possess the relevant moieties. Ultra-large virtual libraries have seen a surge in interest recently due to improvements in computing power. While the traditional library generation approach gravitates to experimentally robust reactions, the invention of new technologies that accelerate invention of new reaction methods necessitates exploration of a wider reaction space, and the role these theoretical reactions play in modulating physicochemical properties.

Naturally, this method is not restricted towards amine–carboxylic acid systems or only their functional groups, α, and β carbons. The same strategies can be applied to other functional group pairs and expanded beyond atoms close to the functional group, though it should be noted that computational time and storage requirements grow exponentially with the number of atoms considered. We are currently preparing a publication that surveys the coupling space between other functional groups such as alcohols, aldehydes, bromides and boronates, exploring the additional chemical space available if coupling conditions were discovered between every combination of these functional groups[54]. A complementary method of extensive enumeration over a few atoms will be enumeration of simpler reactions over an entire molecule, such as C–H oxidation, atom swapping, and atom insertion or deletion, to map the immediate chemical space around a single molecule accessible with other state-of-the-art reaction methodologies[55–60].

2.9 References

(1) Boström, J.; Brown, D. G.; Young, R. J.; Keserü, G. M. Expanding the Medicinal Chemistry Synthetic Toolbox. *Nat. Rev. Drug Discov.* **2018**, *17* (10). https://doi.org/10.1038/nrd.2018.116.

(2) Mahjour, B.; Shen, Y.; Liu, W.; Cernak, T. A Map of the Amine–Carboxylic Acid Coupling System. *Nature* **2020**, *580* (7801). https://doi.org/10.1038/s41586-020-2142-y.

(3) Rains, E. M.; Sloane, N. J. A. On Cayley's Enumeration of Alkanes (or 4-Valent Trees). **2002**.

(4) Henze, H. R.; Blair, C. M. The Number of Isomeric Hydrocarbons of the Methane Series. *J. Am. Chem. Soc.* **1931**, *53* (8). https://doi.org/10.1021/ja01359a034.

(5) Henze, H. R.; Blair, C. M. The Number of Structurally Isomeric Alcohols of the Methanol Series. *J. Am. Chem. Soc.* **1931**, *53* (8). https://doi.org/10.1021/ja01359a027.

(6) Perry, D. The Number of Structural Isomers of Certain Homologs of Methane and Methanol. *J. Am. Chem. Soc.* **1932**, *54* (7). https://doi.org/10.1021/ja01346a035.

(7) Parks, C. A.; Hendrickson, J. B. Enumeration of Monocyclic and Bicyclic Carbon Skeletons. *J. Chem. Inf. Comput. Sci.* **1991**, *31* (2). https://doi.org/10.1021/ci00002a021.

(8) Spialter, Leonard. The Atom Connectivity Matrix (ACM) and Its Characteristic Polynomial (ACMCP): A New Computer-Oriented Chemical Nomenclature. *J. Am. Chem. Soc.* **1963**, *85* (13), 2012–2013. https://doi.org/10.1021/ja00896a022.

(9) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. In *Computers in Chemistry*; Springer-Verlag: Berlin/Heidelberg; pp 19–64. https://doi.org/10.1007/BFb0051317.

(10) Bauer, J. IGOR2: A PC-Program for Generating New Reactions and Molecular Structures. *Tetrahedron Comput. Methodol.* **1989**, *2* (5). https://doi.org/10.1016/0898-5529(89)90034-1.

(11) Ugi, I.; Stein, N.; Knauer, M.; Gruber, B.; Bley, K.; Weidinger, R. New Elements in the Representation of the Logical Structure of Chemistry by Qualitative Mathematical Models and Corresponding Data Structures. In *Computer Chemistry*; Springer Berlin Heidelberg: Berlin, Heidelberg. https://doi.org/10.1007/BFb0111463.

(12) Ugi, I.; Gillespie, P. Representation of Chemical Systems and Interconversions Bybe Matrices and Their Transformation Properties. *Angew. Chem. Int. Ed. Engl.* **1971**, *10* (12). https://doi.org/10.1002/anie.197109141.

(13) Jochum, C.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance(PMCD). *Angew. Chem. Int. Ed. Engl.* **1980**, *19* (7). https://doi.org/10.1002/anie.198004953.

(14) Bauer, J.; Fontain, E.; Forstmeyer, D.; Ugi, I. Interactive Generation of Organic Reactions by IGOR 2 and the PC-Assisted Discovery of a New Reaction. *Tetrahedron Comput. Methodol.* **1988**, *1* (2). https://doi.org/10.1016/0898-5529(88)90017-6.

(15) Estrada, E. Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 3. Molecules Containing Cycles. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (1), 23–27. https://doi.org/10.1021/ci970030u.

(16) Estrada, E. Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (1), 31–33. https://doi.org/10.1021/ci00023a004.

(17) Estrada, E. Generalized Graph Matrix, Graph Geometry, Quantum Chemistry, and Optimal Description of Physicochemical Properties. *J. Phys. Chem. A* **2003**, *107* (38), 7482–7489. https://doi.org/10.1021/jp0346561.

(18) Randić, M. Novel Molecular Descriptor for Structure—Property Studies. *Chem. Phys. Lett.* **1993**, *211* (4–5), 478–483. https://doi.org/10.1016/0009-2614(93)87094-J.

(19) Randi?, M. On Characterization of Three-Dimensional Structures. *Int. J. Quantum Chem.* **1988**, *34* (S15), 201–208. https://doi.org/10.1002/qua.560340718.

(20) Randić, M.; Jerman-Blažić, B.; Trinajstić, N. Development of 3-Dimensional Molecular Descriptors. *Comput. Chem.* **1990**, *14* (3), 237–246. https://doi.org/10.1016/0097-8485(90)80051-3.

(21) Balasubramanian, K. Geometry-Dependent Characteristic Polynomials of Molecular Structures. *Chem. Phys. Lett.* **1990**, *169* (3), 224–228. https://doi.org/10.1016/0009-2614(90)85192-F.

(22) Estrada, E.; Rodríguez-Velázquez, J. A.; Randić, M. Atomic Branching in Molecules. *Int. J. Quantum Chem.* **2006**, *106* (4), 823–832. https://doi.org/10.1002/qua.20850.

(23) Estrada, E.; Hatano, N. Statistical-Mechanical Approach to Subgraph Centrality in Complex Networks. *Chem. Phys. Lett.* **2007**, *439* (1–3), 247–251. https://doi.org/10.1016/j.cplett.2007.03.098.

(24) Estrada, E. Atom–Bond Connectivity and the Energetic of Branched Alkanes. *Chem. Phys. Lett.* **2008**, *463* (4–6), 422–425. https://doi.org/10.1016/j.cplett.2008.08.074.

(25) Chen, X. On ABC Eigenvalues and ABC Energy. *Linear Algebra Its Appl.* **2018**, *544*, 141–157. https://doi.org/10.1016/j.laa.2018.01.011.

(26) Hosseini, S. A.; Mohar, B.; Ahmadi, M. B. The Evolution of the Structure of ABC-Minimal Trees. *J. Comb. Theory Ser. B* **2022**, *152*, 415–452. https://doi.org/10.1016/j.jctb.2021.07.001.

(27) Janežič, D.; Miličević, A.; Nikolić, S.; Trinajstić, N. *Graph-Theoretical Matrices in Chemistry*; CRC Press: Oxford, 2017.

(28) Fujita, S. Description of Organic Reactions Based on Imaginary Transition Structures. 1. Introduction of New Concepts. *J. Chem. Inf. Comput. Sci.* **1986**, *26* (4), 205–212. https://doi.org/10.1021/ci00052a009.

(29) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57* (8). https://doi.org/10.1021/acs.jcim.6b00601.

(30) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2). https://doi.org/10.1039/C8SC04228D.

(31) Zhao, Q.; Savoie, B. M. Simultaneously Improving Reaction Coverage and Computational Cost in Automated Reaction Prediction Tasks. *Nat. Comput. Sci.* **2021**, *1* (7). https://doi.org/10.1038/s43588-021-00101-3.

(32) Pollock, S. N.; Coutsias, E. A.; Wester, M. J.; Oprea, T. I. Scaffold Topologies. 1. Exhaustive Enumeration up to Eight Rings. *J. Chem. Inf. Model.* **2008**, *48* (7). https://doi.org/10.1021/ci7003412.

(33) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3). https://doi.org/10.1021/ar500432k.

(34) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55* (1). https://doi.org/10.1021/ci5006614.

(35) Zhang, R.; Mahjour, B.; Outlaw, A.; McGrath, A.; Hopper, T.; Kelley, B.; Walters, W. P.; Cernak, T. Exploring the Combinatorial Explosion of Amine–Acid Reaction Space via Graph Editing. *Commun Chem.,* under review.

(36) *RDKit: Open-source cheminformatics*.

(37) 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). https://doi.org/10.1021/ci025599w.

(38) Wishart, D. S. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34* (90001). https://doi.org/10.1093/nar/gkj067.

(39) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**.

(40) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113. https://doi.org/10.1021/c160017a018.

(41) Sotnik, S. O.; Subota, A. I.; Kliuchynskyi, A. Y.; Yehorov, D. V.; Lytvynenko, A. S.; Rozhenko, A. B.; Kolotilov, S. V.; Ryabukhin, S. V.; Volochnyuk, D. M. Cu-Catalyzed Pyridine Synthesis via Oxidative Annulation of Cyclic Ketones with Propargylamine. *J. Org. Chem.* **2021**, *86* (11). https://doi.org/10.1021/acs.joc.0c03038.

(42) Shabalin, D. A. Recent Advances and Future Challenges in the Synthesis of 2,4,6-Triarylpyridines. *Org. Biomol. Chem.* **2021**, *19* (38). https://doi.org/10.1039/D1OB01310F.

(43) Zhang, Z.; Cernak, T. The Formal Cross-Coupling of Amines and Carboxylic Acids to Form Sp ^{3} –Sp ^{3} Carbon–Carbon Bonds. *Angew. Chem. Int. Ed.* **2021**, *60* (52), 27293–27298. https://doi.org/10.1002/anie.202112454.

(44) Shen, Y.; Mahjour, B.; Cernak, T. Development of Copper-Catalyzed Deaminative Esterification Using High-Throughput Experimentation. *Commun. Chem.* **2022**, *5* (1), 83. https://doi.org/10.1038/s42004-022-00698-0.

(45) Douthwaite, J. L.; Zhao, R.; Shim, E.; Mahjour, B.; Zimmerman, P. M.; Cernak, T. Formal Cross-Coupling of Amines and Carboxylic Acids to Form Sp3–Sp2 Carbon–Carbon Bonds. *J. Am. Chem. Soc.* **2023**, *145* (20), 10930–10937. https://doi.org/10.1021/jacs.2c11563.

(46) McGrath, A.; Zhang, R.; Shafiq, K.; Cernak, T. Repurposing Amine and Carboxylic Acid Building Blocks with an Automatable Esterification Reaction. *Chem. Commun.* **2023**, *59* (8), 1026–1029. https://doi.org/10.1039/D2CC05670D.

(47) Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W. D.; Tomkinson, N. P. Designing Novel Building Blocks Is an Overlooked Strategy to Improve Compound Quality. *Drug Discov. Today* **2015**, *20* (1), 11–17. https://doi.org/10.1016/j.drudis.2014.09.023.

(48) Grygorenko, O. O.; Volochnyuk, D. M.; Vashchenko, B. V. Emerging Building Blocks for Medicinal Chemistry: Recent Synthetic Advances. *Eur. J. Org. Chem.* **2021**, *2021* (47), 6478–6510. https://doi.org/10.1002/ejoc.202100857.

(49) Pennington, L. D.; Aquila, B. M.; Choi, Y.; Valiulin, R. A.; Muegge, I. Positional Analogue Scanning: An Effective Strategy for Multiparameter Optimization in Drug Design. *J. Med. Chem.* **2020**, *63* (17), 8956–8976. https://doi.org/10.1021/acs.jmedchem.9b02092.

(50) Helal, C. J.; Bundesmann, M.; Hammond, S.; Holmstrom, M.; Klug-McLeod, J.; Lefker, B. A.; McLeod, D.; Subramanyam, C.; Zakaryants, O.; Sakata, S. Quick Building Blocks (QBB): An Innovative and Efficient Business Model To Speed Medicinal Chemistry Analog Synthesis. *ACS Med. Chem. Lett.* **2019**, *10* (8), 1104–1109. https://doi.org/10.1021/acsmedchemlett.9b00205.

(51) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51* (D1), D1373–D1380. https://doi.org/10.1093/nar/gkac956.

(52) Xiao, Q.; Wang, L.; Supekar, S.; Shen, T.; Liu, H.; Ye, F.; Huang, J.; Fan, H.; Wei, Z.; Zhang, C. Structure of Human Steroid 5α-Reductase 2 with the Anti-Androgen Drug Finasteride. *Nat. Commun.* **2020**, *11* (1), 5430. https://doi.org/10.1038/s41467-020-19249-z.

(53) Tsitsanou, K. E.; Thireou, T.; Drakou, C. E.; Koussis, K.; Keramioti, M. V.; Leonidas, D. D.; Eliopoulos, E.; Iatrou, K.; Zographos, S. E. Anopheles Gambiae Odorant Binding Protein Crystal Complex with the Synthetic Repellent DEET: Implications for Structure-Based Design of Novel Mosquito Repellents. *Cell. Mol. Life Sci.* **2012**, *69* (2), 283–297. https://doi.org/10.1007/s00018-011-0745-z.

(54) Mahjour, B.; Zhang, R.; Outlaw, A.; Zhang, X.; Harmata, A.; Cernak, T. Highlighting Opportunities for Reaction Invention Through Analysis of Commercially Available Chemical Building Blocks. *Org. Lett.,* under review.

(55) Woo, J.; Stein, C.; Christian, A. H.; Levin, M. D. Carbon-to-Nitrogen Single-Atom Transmutation of Azaarenes. *Nature* **2023**, *623* (7985), 77–82. https://doi.org/10.1038/s41586-023-06613-4.

(56) Kennedy, S. H.; Dherange, B. D.; Berger, K. J.; Levin, M. D. Skeletal Editing through Direct Nitrogen Deletion of Secondary Amines. *Nature* **2021**, *593* (7858), 223–227. https://doi.org/10.1038/s41586-021-03448-9.

(57) Lyu, H.; Kevlishvili, I.; Yu, X.; Liu, P.; Dong, G. Boron Insertion into Alkyl Ether Bonds via Zinc/Nickel Tandem Catalysis. *Science* **2021**, *372* (6538), 175–182. https://doi.org/10.1126/science.abg5526.

(58) Gensch, T.; Hopkinson, M. N.; Glorius, F.; Wencel-Delord, J. Mild Metal-Catalyzed C–H Activation: Examples and Concepts. *Chem. Soc. Rev.* **2016**, *45* (10), 2900–2936. https://doi.org/10.1039/C6CS00075D.

(59) Hartwig, J. F. Catalyst-Controlled Site-Selective Bond Activation. *Acc. Chem. Res.* **2017**, *50* (3), 549–555. https://doi.org/10.1021/acs.accounts.6b00546.

(60) Rogge, T.; Kaplaneris, N.; Chatani, N.; Kim, J.; Chang, S.; Punji, B.; Schafer, L. L.; Musaev, D. G.; Wencel-Delord, J.; Roberts, C. A.; Sarpong, R.; Wilson, Z. E.; Brimble, M. A.; Johansson, M. J.; Ackermann, L. C–H Activation. *Nat. Rev. Methods Primer* **2021**, *1* (1), 43. https://doi.org/10.1038/s43586-021-00041-2.

Chapter 3 Applying Matrix Encoding of Reactions to Analysis of Organic Synthesis Pathways

3.1 Introduction and background

The utility of encoding organic reactions as matrices can be expanded beyond single-step reaction enumeration, and applied towards analysis of multi-step synthetic pathways. When presented with a total synthesis route, experienced chemists can readily apply knowledge and intuition to evaluate efficiencies of each step. For example, highly effective reactions are those that form multiple bonds, such as cycloaddition and cascade reactions[1], or those that unite multiple building blocks into one[2]. On the other hand, reactions that manipulate protecting groups[3], undergo unnecessary redox operations[4], or perform multiple functional group interconversions[5] are judged unfavorably.

Several works have aimed to quantify this intuition and express them as visualizations of synthetic pathways. Hendrickson, through graph encoding of intermediates, depicts a synthetic route as a traversal along two axes[6], one representing the number of building blocks and the other ring count. While being a unique approach that highlights convergent and ring-forming steps, each route can overlap with itself multiple times, causing clutter when visualizing long routes. Steps that do not modify ring or building block count also cannot be represented. Modern approaches broadly fall into two categories, using either a simple graph encoding intermediates as nodes and steps as weighted edges based on properties such as reaction yield[7] or type[8], or using a line graph plotting computed descriptors of intermediate against synthesis progress. The choices for descriptors frequently include molecular complexity[9–11], similarity to the target[12], or physicochemical properties such as FSP3 and MW[13].

3.2 Strategy of synthetic route analysis through graph edit distance plots.

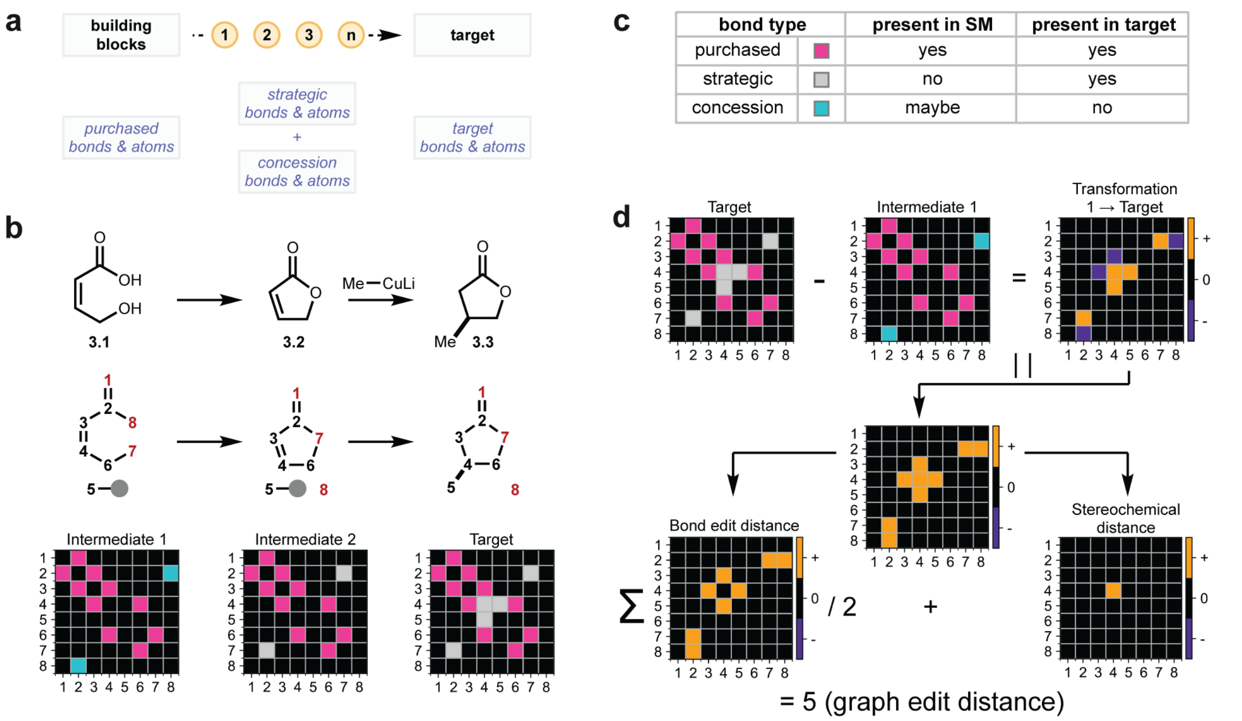
3.2.1 Measurement of relative step impact

The application of matrix-encoded transformations herein aims to address limitations we observe in current synthetic route visualization algorithms. The method of simple graphs places equal weight on all steps that fall under a certain reaction class, and hence does not provide insight on their relative impacts. Plots of molecular properties against synthesis progress can highlight impactful steps, but the complexity of algorithms employed to compute these properties result, to varying extents, to a "black box" visualization – the graphs can show which steps are impactful, but not why they are rated as such.

We propose graph economy to be an effective measure of both step impact and overall route efficiency. A synthetic pathway can be conceptualized as a sequence of reactions that convert one or more building blocks, comprising of purchased bonds and atoms, into the target molecule's bonds and atoms. Throughout the synthesis, strategic bonds, concession bonds, and concession atoms will be added and removed from these building blocks (Fig. 3.1a).

To conduct graph economy analysis of a synthetic route, the target molecule, starting materials and all intermediates are converted into matrix representations, following the workflow in Chapter 2.3 (Fig. 3.1b), with some additional considerations: the presence of a stereocenter is encoded by entries on the diagonal, atom indices must be consistent throughout the route, and all atoms that appear at least once on an intermediate must receive an index. For example, in the simple route depicted in Fig. 3.1b, both oxygen #8, which is removed after the first step, and carbon #5, which is only added in the second step, receive indices. In contrast, atoms and groups that do not appear as part of the main intermediate at any point, such as the CuLi group, are omitted.

The identities of each bond are evaluated on its presence in the starting materials and products (Fig. 3.1c). A purchased bond is present in both the target and starting materials, being desired bond that is present at the outset. A strategic bond is present in the target but absent in the starting materials, and needs to be formed during the synthesis. Lastly, a concession bond is absent in the target, and may or may not be present in the starting material. These are bonds need not be theoretically formed, but appear transiently in the route. One common example is a bond connecting the target scaffold to a protecting group, which serves to prevent undesired reactions in subsequent steps.



bond type		present in SM	present in target
purchased	🟥	yes	yes
strategic	⬜	no	yes
concession	🟦	maybe	no

Figure 3.1. Matrix-encoded transformations in analysis of total synthesis routes. **a.** A high-level view of an organic synthesis route, where building blocks are converted to the target via a sequence of reactions. **b.** A simple synthetic route to demonstrate matrix encoding of multiple reactions in sequence, with matrix encoding of each intermediate to demonstrate different bond types. **3.1** first cyclizes to lactone **3.2**, followed by a stereoselective conjugate addition to **3.3**. **c.** Classifications of bond types used in matrix representations. **d.** Visualization of how graph edit distance (GED) between two matrix-encoded molecular systems is computed. Transformation matrices only show whether an entry is positive (yellow) or negative (purple) in the visualization, but their numerical entries are tracked in practice.

With matrix representations in hand, the transformation matrices between each intermediate and target are computed by subtracting each intermediate matrix from the target matrix (Fig. 3.1d). This matrix encodes all bond and stereocenter edits that the intermediate must undergo to arrive at the target structure. Next, the absolute value of the transformation matrix is taken, then the matrix split into the bond and stereocenter differences. Bond edit distance is taken by summing up all entries in the bond difference matrix, then dividing by 2, as the symmetry means that each bond edit is counted twice. Each stereocenter is only encoded once, hence the stereocenter difference is simply the sum of all diagonal elements. Finally, the bond and stereocenter edit distances are summed to obtain the intermediate's graph edit distance (GED) to the target.

Since GED measures the degree of similarity between two molecular systems, the plot of its value against synthesis progress generally shows a downwards trend, reflecting how intermediates are becoming increasingly similar to the target. The relative impact of each synthetic step can then be compared through their relative slopes – high-impact steps, which do more work in bringing the intermediate closer to the target, appear as steep negative slopes, while lower-impact steps have a less negative, or even positive slope.

Fig. 3.2a shows the visualization of Heathcock and Stafford's synthesis of (–)-secodaphniphylline[14] (**3.4**) in 9 steps from various building blocks (**3.5**–**3.8**). From examining the plot of GED against synthetic intermediates, step 6 is readily apparent as the highest impact step, (Fig. 3.2b) being the remarkable tetracyclization reaction that forms several target bonds while eliminating two concession alcohol groups. In contrast, lower-impact steps, such as the lactonization step 4 (Fig. 3.2c), involve fewer bond edits.

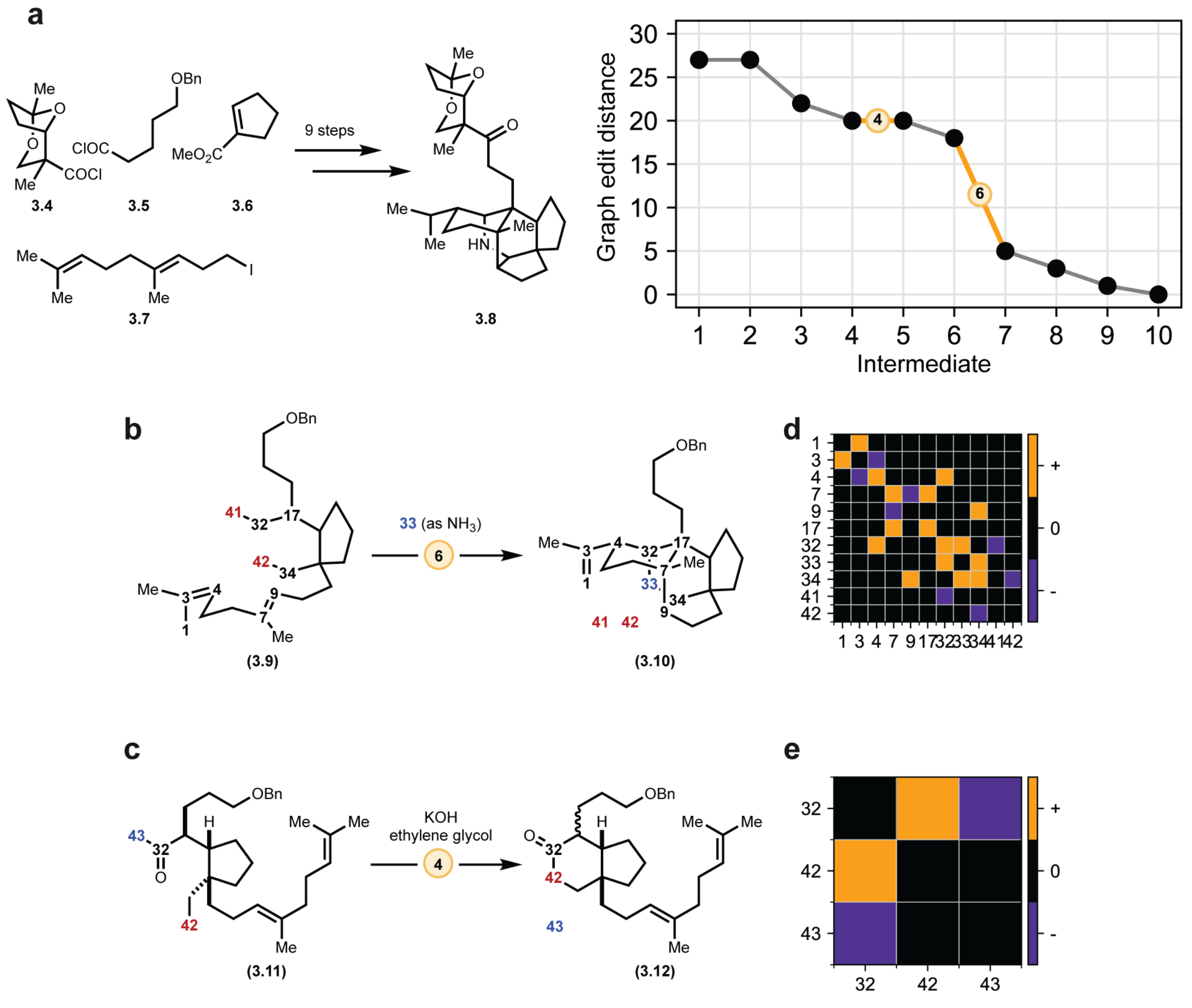


Figure 3.2. Graph edit distance plots enable identification of key synthetic steps. **a.** (–)-secodaphniphylline (**3.4**) is synthesized from building blocks **3.5**-**3.8** in nine steps, as visualized in the GED plot. **b.** Key cascade step as identified by GED analysis. Indices are provided at atoms where transformations occur, accompanied by its transformation matrix. **c.** Example of a low-impact step, accompanied by its transformation matrix. Indices are provided at atoms where transformations occur. Black indices = carbon, blue indices = nitrogen, red indices = oxygen. Structures of all intermediates can be found in Appendix B.

Due to the matrix encoding of all intermediates, the transformation matrix for any individual step can be readily extracted to display the bonds which are changing in that step. This can provide insight into the reason behind a step's calculated impact. Visualization of the transformation matrix of step 4 (Fig. 3.2d) reveals that numerous bonds and stereocenters are being formed, whereas there is much less activity in step 4 (Fig. 3.2e). Furthermore, even though the C32–N43 concession bond is broken in step 4, the created 32–42 bond is also a concession bond, hence there is no overall change in GED to the target.

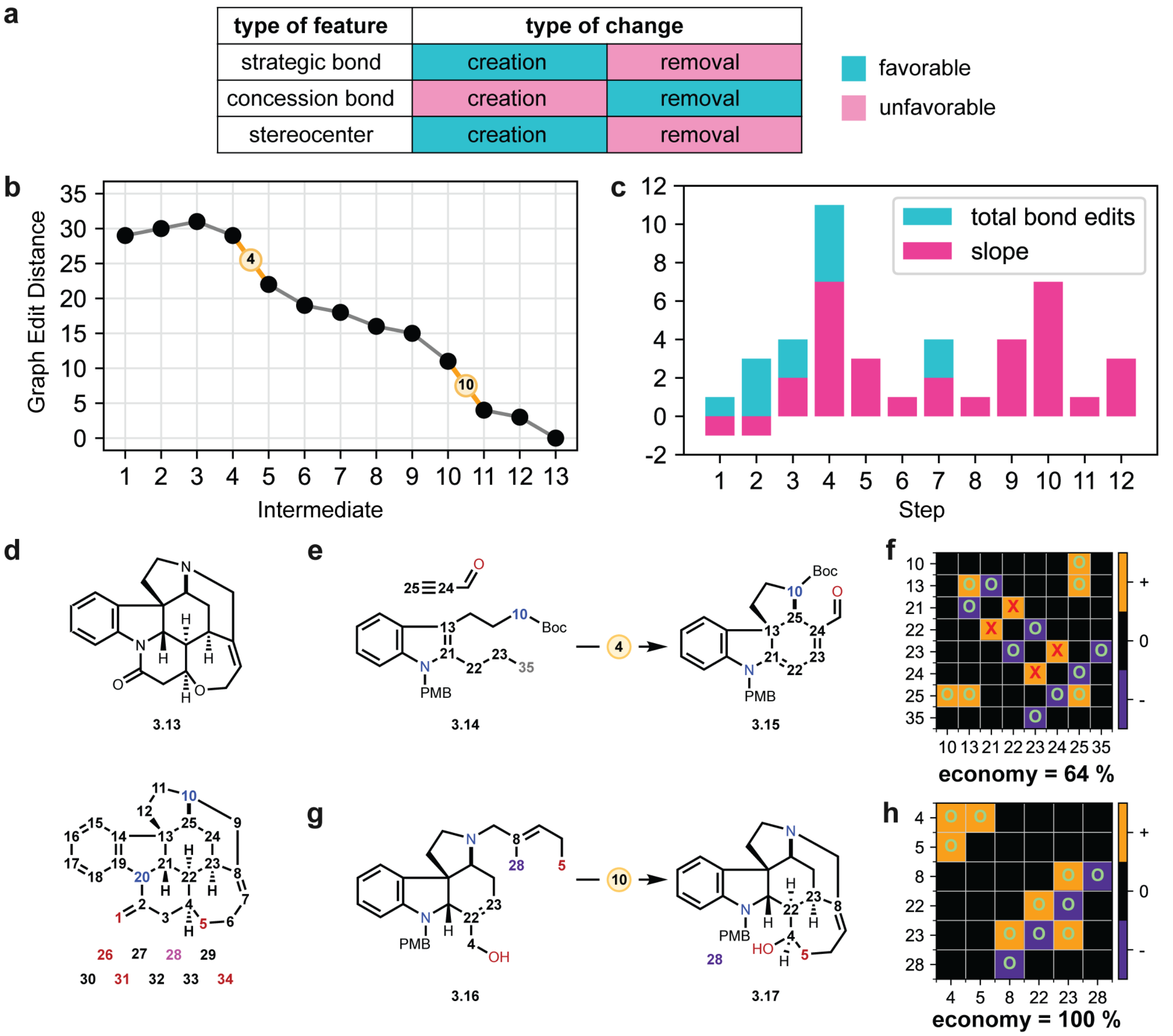
3.2.2 Calculation of graph economy in individual synthetic steps.

As the previous chapter demonstrates, not all transformations are conducive to reducing the distance between intermediate and product. For example, formation of a strategic bond will reduce the GED, but formation of a concession bond will result in the opposite. Fig. 3.3a describes the effects of each change in bond types and stereocenters – creation of strategic bonds and stereocenters are favorable to synthesis progress, while their removal is unfavorable. Concession bonds, being absent in the target, have the opposite effect. The graph economy of an individual step can be calculated by dividing the number of favorable changes by the total number of bond and stereocenter changes made in that step. A step has 100% graph economy if all changes are favorable, and the percentage decreases if unfavorable bond changes are occurring simultaneously.

In the synthesis of strychnine (**3.13**) by MacMillan and coworkers[15], a GED plot highlights steps 4 and 10 as the highest-impact steps (Fig. 3.3b). However, a bar plot overlaying total bond edits and negative slope of each step reveals that the slope of step 4 was lower than its bond edit count, whereas these amounts are equal in step 10, as visualized by absence of the blue bar (Fig. 3.3c). This difference in bond economy can be analyzed by extracting the full set of target atom indices (Fig. 3.3d) to cross-examine against the transformation matrix of each step. As the presence of all bonds throughout the synthesis have been encoded, each entry in the transformation matrices can be annotated based on their impact on the reduction of GED. An entry annotated with an "o" has a favorable impact, while an entry with an "x" is unfavorable.

Both high-impact steps in the strychnine synthesis are identified as cascade reactions, forming many strategic bonds and stereocenters. However, visualization of step 4's transformation matrix (Fig. 3.3e) reveals the creation of two concession bonds, one between C21 and C22, and the other between C23 and C24. In contrast, the second cascade reaction, step 10 (Fig. 3.3f), has

100% graph economy as all elements of its transformation matrix contributed to the lowering of GED between intermediate and target.



type of feature	type of change	
strategic bond	creation	removal
concession bond	creation	removal
stereocenter	creation	removal

🟦 favorable
🟪 unfavorable

Figure 3.3. Impact of manipulating different bond types on synthetic progress. **a.** The impact of strategic bonds, concession bonds, and stereocenters on the reduction of GED between intermediate and target. **b.** GED plot of the synthesis of strychnine by MacMillan and coworkers. **c.** Graph economy plot of the same synthesis. The maximum height of each bar depicts the total number of bond edits in that step, while the height of the pink bars depicts the negative slope of each step. The degree of coverage by the pink bars in each step represents its graph economy, with 100% graph economy if there is no blue visible. **d.** Structure of the target strychnine with indices of all atoms. **e.** A cascade reaction evaluated to be of high impact, but contains inefficiencies in graph economy, denoted by "x" markings in its transformation matrix. **f.** A high-impact cascade reaction that has 100% graph economy. Indices are provided at atoms where transformations occur. Black indices = carbon, blue indices = nitrogen, red indices = oxygen.

3.3 Application towards total synthesis of stemoamide

The method of GED analysis was applied towards complementing our group's computer-aided total synthesis of stemoamide (**3.18,** Fig. 3.4a), a compound present in *Stemonaceae* plant roots which are used in traditional Chinese medicine for treatment of respiratory illnesses[16]. In addition to the synthetic challenges presented by its fused-ring structure and four stereocenters, stemoamide and its congeners, such as ethylstemoamide (**3.19**), are also intermediates in the synthesis of higher stemona alkaloids that exhibit exciting bioactivities, such as stemonine[17] (**3.20**), a remarkable inhibitor of lung fibrosis *in vitro* and *in vivo*, and sessilifoliamide A[18] (**3.21**), an *in vitro* inhibitor of nitrous oxide release. Efficient synthetic routes developed towards stemoamide will also bolster efforts towards these higher stemona alkaloids.

To explore the potential of computer-aided synthesis planning (CASP) in total synthesis of complex alkaloids, we first performed automated retrosynthesis of stemoamide with the CASP software SYNTHIA™. Given a set of scoring criteria, the software will perform iterative disconnection using encoded reaction rules, until all building blocks are either commercially available or have been previously reported in literature (Fig. 3.4b). A surprising outcome was the presence of the Mannich reaction in all predicted routes, as it has not been featured in any previous syntheses of this target. One example route is shown in Fig. 3.4c: after hydroxymethylation of **3.22** to **3.23**, an asymmetric organocatalyzed Mannich is proposed to unite it with **3.24** and **3.25** to form **3.26**, with subsequent allylation and functional group interconversions to form the remaining rings and arrive at (-)-**3.18.**

At the time of this study, the resultant seven-step synthesis was on par with, but not shorter than the shortest reported asymmetric route[17]. To further reduce step count, we employed GED analysis with the goal of extracting high-impact steps to incorporate into our own experimental synthesis route, while minimizing the number of lower-impact steps. In accordance with CASP proposals, the Mannich coupling was underscored as the step with highest impact, furthering the motivation to incorporate this reaction into our strategy.

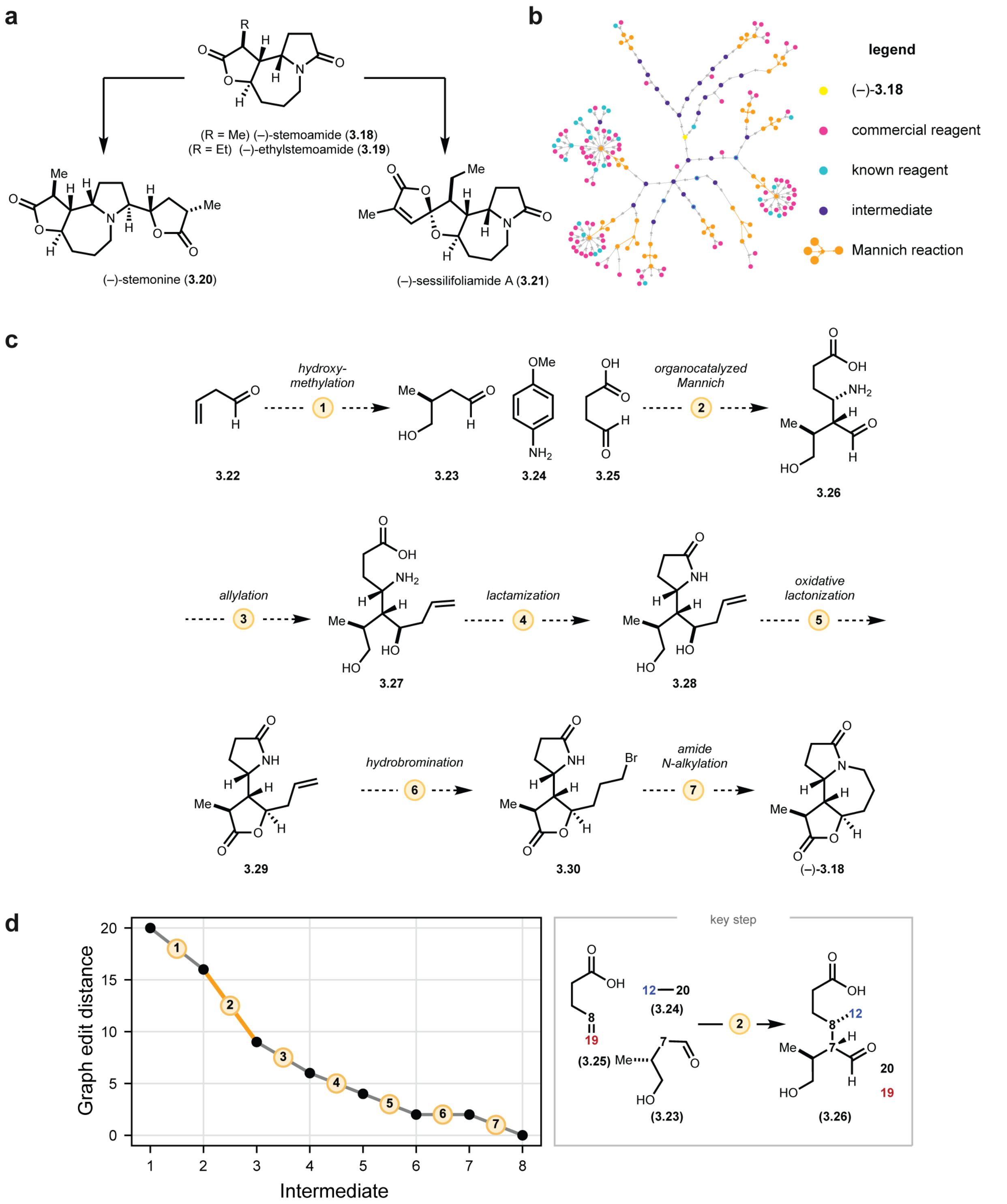


Figure 3.4. Applying GED plots in analysis of a computer-proposed synthesis route. **a.** Stemoamide (**3.18**), the target of our group's total synthesis work, along with analog **3.19**, is an important intermediate towards the synthesis of higher stemona alkaloids such as **3.20** and **3.21**. **b.** Network of 50 SYNTHIA™-predicted routes to (–)-**3.18**. The Mannich reaction, represented by a cluster of four orange dots in each route, appears as a consistent disconnection. **c.** A synthetic route towards (–)-**3.18**, proposed by SYNTHIA™. **d.** GED plot of the route in **c**, accompanied by the identified key step.

Through coupling the findings of GED analysis with manual examination of the computer-proposed route, we completed a six-step asymmetric synthesis of **3.18**[19]. An optimization of redox economy eliminated the first hydroxylation step (Fig. 3.5a, step 1), resulting in a self-Mannich reaction between two equivalents of **3.31**[20,21], a commercially available compound. Two other lower-impact steps, allylation and oxidative lactonization (Fig. 3.5a, steps 3 and 5), were further absorbed into this synthetic step. Attempts to perform subsequent hydrobromination were met with difficulty, hence the single computer-proposed step was split into two – first converting alkene **3.35** into alcohol **3.36**, followed by bromination and *in-situ* removal of the *p*-methoxyphenyl group to arrive at bromide **3.37**. Closure of the final seven-membered ring[22], and finally methylation[17] affords the target (+)-**3.18**, achieving the shortest asymmetric synthesis of this target at the time of discovery.

GED analysis of this experimental synthesis highlights the impacts of our modifications to the computer-proposed route, contrasting the steep negative slope of step 1 (Fig. 3.5b) with the lower impact of the subsequent functional group interconversions, especially step 3 with zero slope, being an added concession to enable bromination of **3.35**.

The method of GED analysis allows multiple synthetic routes towards the same target to be overlaid, and their relative performances examined. By overlaying the two routes to **3.18**, the avenue through which the experimental route improves on the computer-proposed method can be identified. Our experimental route, through merging two lower-impact functional group interconversions with the high-impact Mannich coupling, produces a steeper slope in its key step (Fig. 3.6a). This is affirmed through comparison of the key steps' transformation matrices (Fig. 3.6b). While the computer-proposed key step already presents a 100% bond economy, our method enables the forging of one additional stereocenter and several other strategic bonds, without employing any concession bonds (Fig. 3.6c-d).

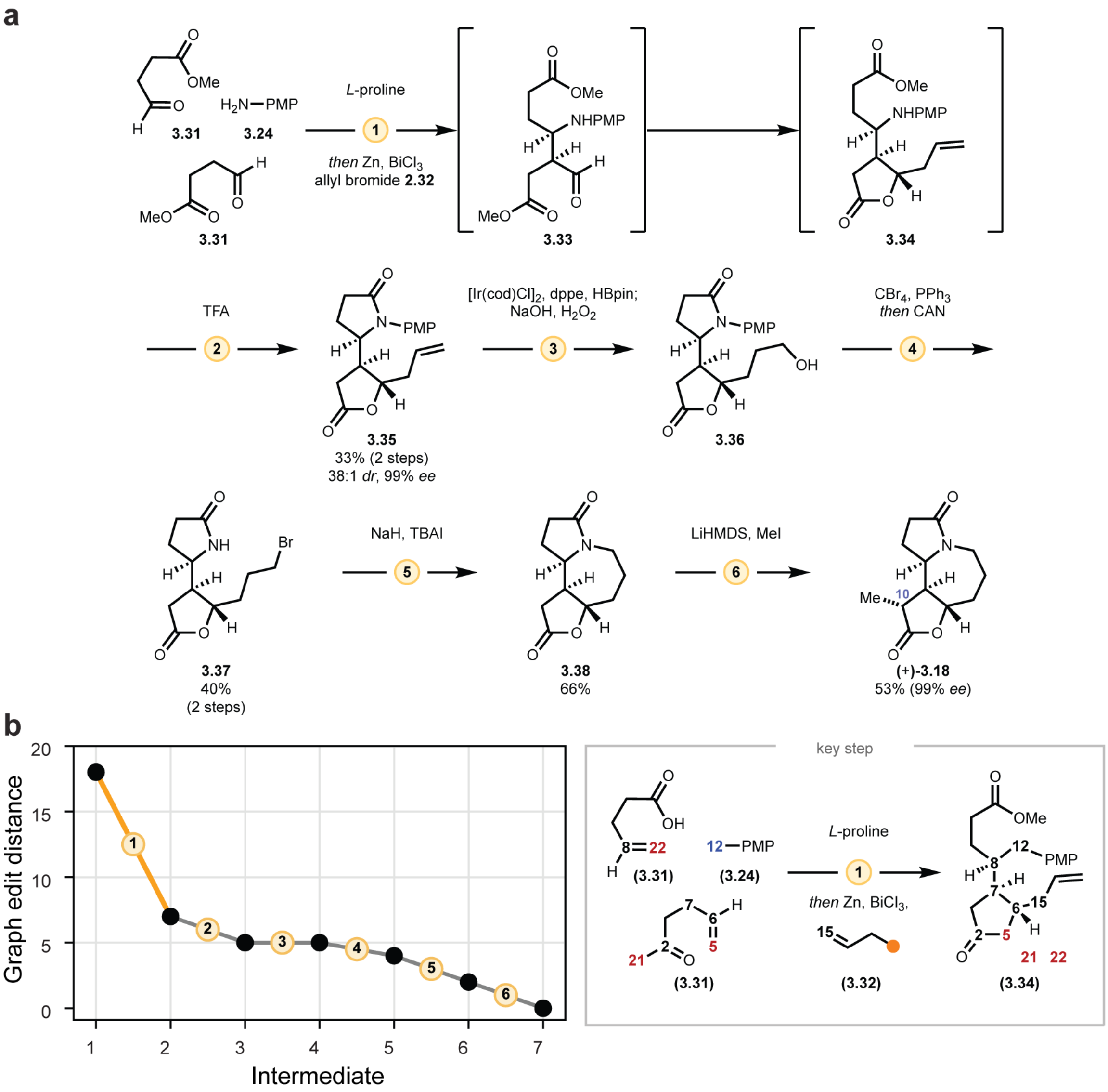


Figure 3.5. Experimental total synthesis of (+)-**3.18** in six steps. TFA = trifluoroacetic acid, dppe =1,2-Bis(diphenylphosphino)ethane; CAN = ceric ammonium nitrate; TBA = tetra-n-butylammonium; HMDS = hexamethyldisilazide. **b.** GED plot of the synthetic route, with the key step highlighted.

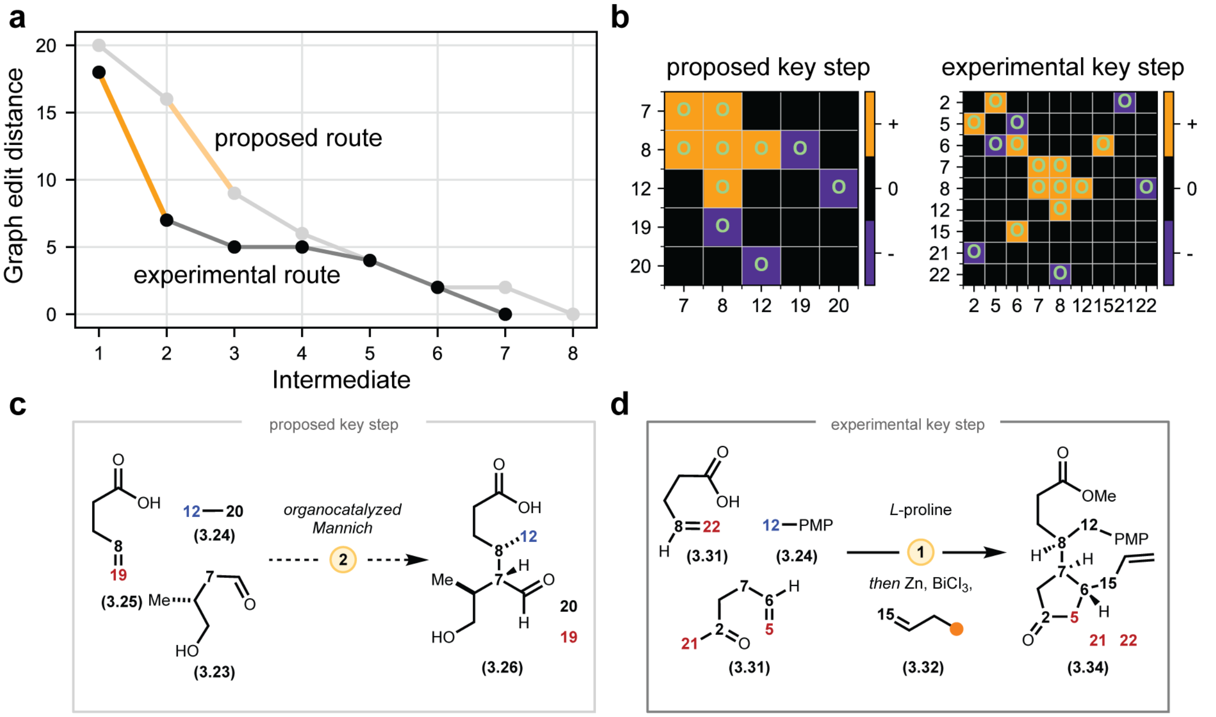


Figure 3.6. Overlay of multiple GED plots enable comparision between routes. **a.** GED plot of both the computer-proposed route (light grey) and experimental route (black) towards (+)-**3.18,** with key steps highlighted in orange. **b.** Transformation matrices of the key step in each route, with each entry annotated by their impact towards reduction in GED. **c.** Scheme of the computer-proposed key step, with reacting atoms annotated with their indices. **d.** Scheme of the experimental key step, with reacting atoms annotated with their indices. Structure numbers are provided in parentheses. Indices are provided at atoms where transformations occur. Black indices = carbon, blue indices = nitrogen, red indices = oxygen.

Seeking to further improve our synthesis of **3.18**, we generated more routes to this target, aiming to repurpose multiple computer-proposed key steps into a single synthesis. Two different search strategies were employed – exclusion of the Mannich reaction as a retrosynthetic disconnection reveals a Michael addition–alkylation key step (Fig. 3.7a), while a search starting from the penultimate intermediate **3.38** highlights lactam formation through Schmidt rearrangement of a cyclobutanone intermediate (Fig. 3.7b). Through adapting both aforementioned key steps into a single route, we achieved a short synthesis of (-)-**3.18** in only three steps (Fig. 3.7c). Allylation of **3.43** produces **3.44**, containing the same lactone motif as computer-proposed **3.39**, which undergoes Michael addition with the Enders (S)-1-amino-2-methoxymethylpyrrolidine (SAMP) hydrazone of cyclobutanone (**3.45**). Methylation with methyl iodide followed by acidic quench affords **3.46**, and finally *anti*-Markovnikov hydroazidation followed by an intramolecular Schmidt-Aubé rearrangement leads to the target (-)-**3.18** in half our previous step count. Furthermore, we have improved upon the graph economy of both computer-proposed routes from 56% and 70% to 100% economy in our final experimental route, through extracting steps with full individual graph economy from CASP (Fig. 3.7e).

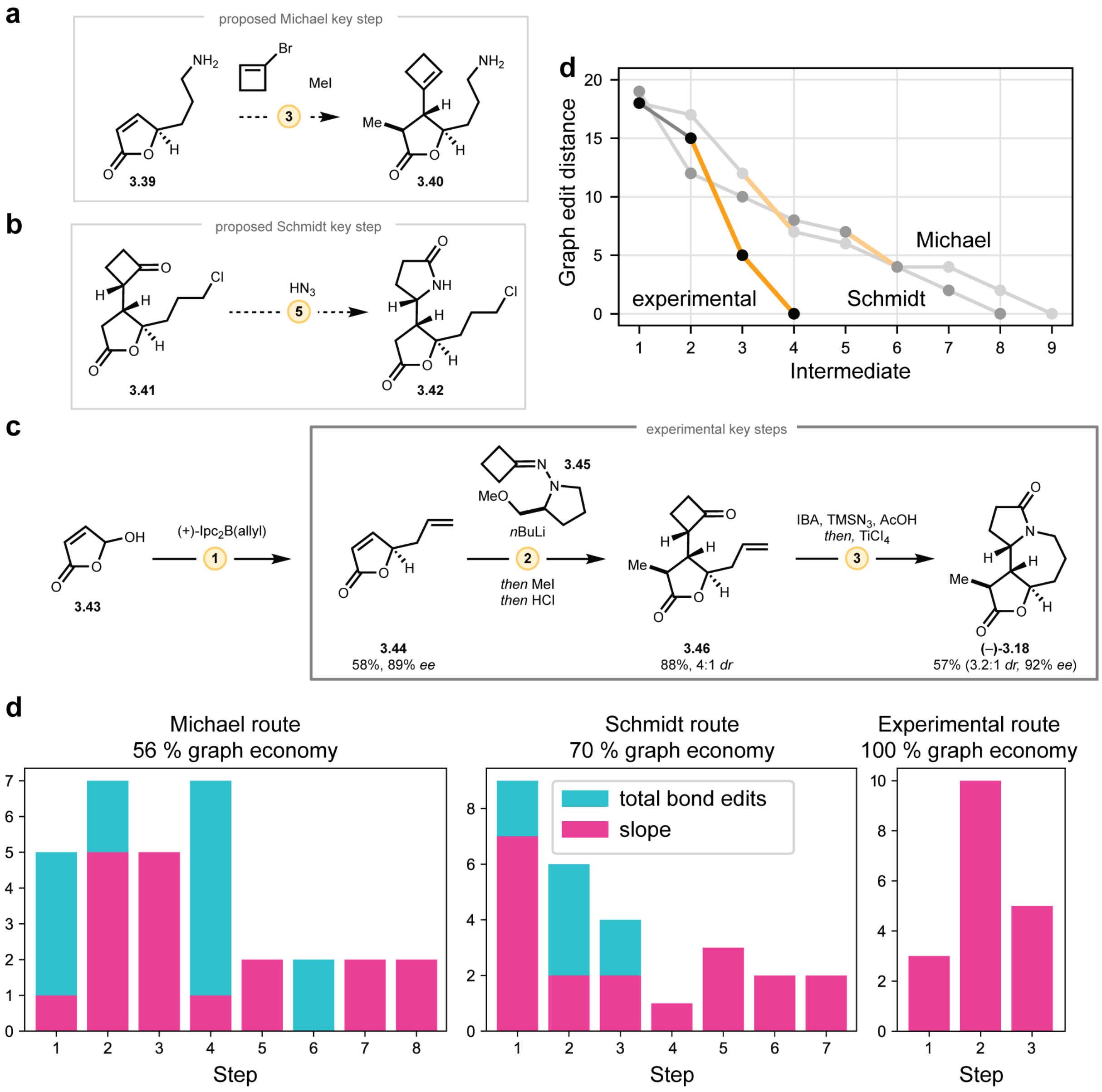


Figure 3.7. Combining multiple key steps into a single route. **a.** A computer-proposed Michael addition-alkylation key step. **b.** A computer-proposed Schmidt rearrangement key step. **c.** Our experimental three-step synthesis of (–)-**3.18** achieved by adapting key steps in both **a** and **b**. (+)-Ipc2B(allyl) = (+)-B-Allyldiisopinocampheylborane; IBA = iodosobenzoic acid; TMS = trimethylsilyl **d.** GED plots of both computer-proposed routes and the final experimental route. The steep first step of the Schmidt route was not selected due to an incorrect selectivity. **e.** Graph economy plots of all three routes. Structures of all intermediates in **a** and **b** can be found under Appendix B.

As of publication, we have developed the two shortest asymmetric syntheses of **3.18**, through the tandem effort of the CASP program SYNTHIA™ in generating large amounts of proposed routes, GED analysis in highlighting key steps, and human knowledge to adapt these key steps into experiment.

3.4 Incorporation of the Maximum Common Substructure (MCS) difference as an additional metric

In comparing our synthetic route visualization method with other contemporary methods, we observed that several methods call attention to convergent steps[6–8], which are reactions that couple two fragments, usually of equal molecular complexity. Convergent routes are highly sought after in total synthesis design[23–25], as preparation of building blocks separately reduces the amount of functional group compatibility considerations per step, resulting in lower step count and consequently improved overall yield[26,27].

On the theoretical level, a convergent step should have favorable graph economy, as multiple target atoms are assembled into one structure. However, examination of GED analyses of various reported routes in synthetic literature reveals that convergent steps are not necessarily computed as high-impact steps. In the synthesis of welwitindolinone A by Baran and Richter[28] (Fig. 3.8a), GED analysis presents the highest-impact step as the last oxidative ring contraction (Fig. 3.8b, grey dots), but assigns a low score to step 1, despite it being a significant convergent coupling with indole. As the entire molecular system is encoded as a single matrix, the impact of each step is only calculated based on bonds and stereocenters formed, without considering whether these bonds reduce the number of connected fragments.

Aiming to place more importance on convergent steps, we added an additional metric, the maximum common substructure (MCS) difference to the GED of each intermediate. The MCS between an intermediate and target is the largest interconnected motif in the intermediate that has identical atom identities and valencies as the target, that is, it can be accurately overlaid onto the target structure. Fig. 3.8a visualizes how the MCS, represented by the purple area in each intermediate, changes with synthetic progress. In contrast to GED, the number of heavy atoms in each intermediate's MCS trend upwards with synthesis progress. To harmonize the two metrics, we chose to invert the MCS size into MCS difference by subtracting the heavy atom count in each intermediate's MCS from that of the target. A plot of MCS difference against synthetic intermediates is shown in Fig. 3.8b in purple dots, demonstrating how step 1, the only convergent step which couples of **3.47** with indole (**3.48**), is now highlighted as the most impactful step. By scaling the MCS difference of each intermediate by half, then adding the result to GED, we were able to achieve an improved step impact metric that highlights both convergent steps, and those that make many favorable bond changes.

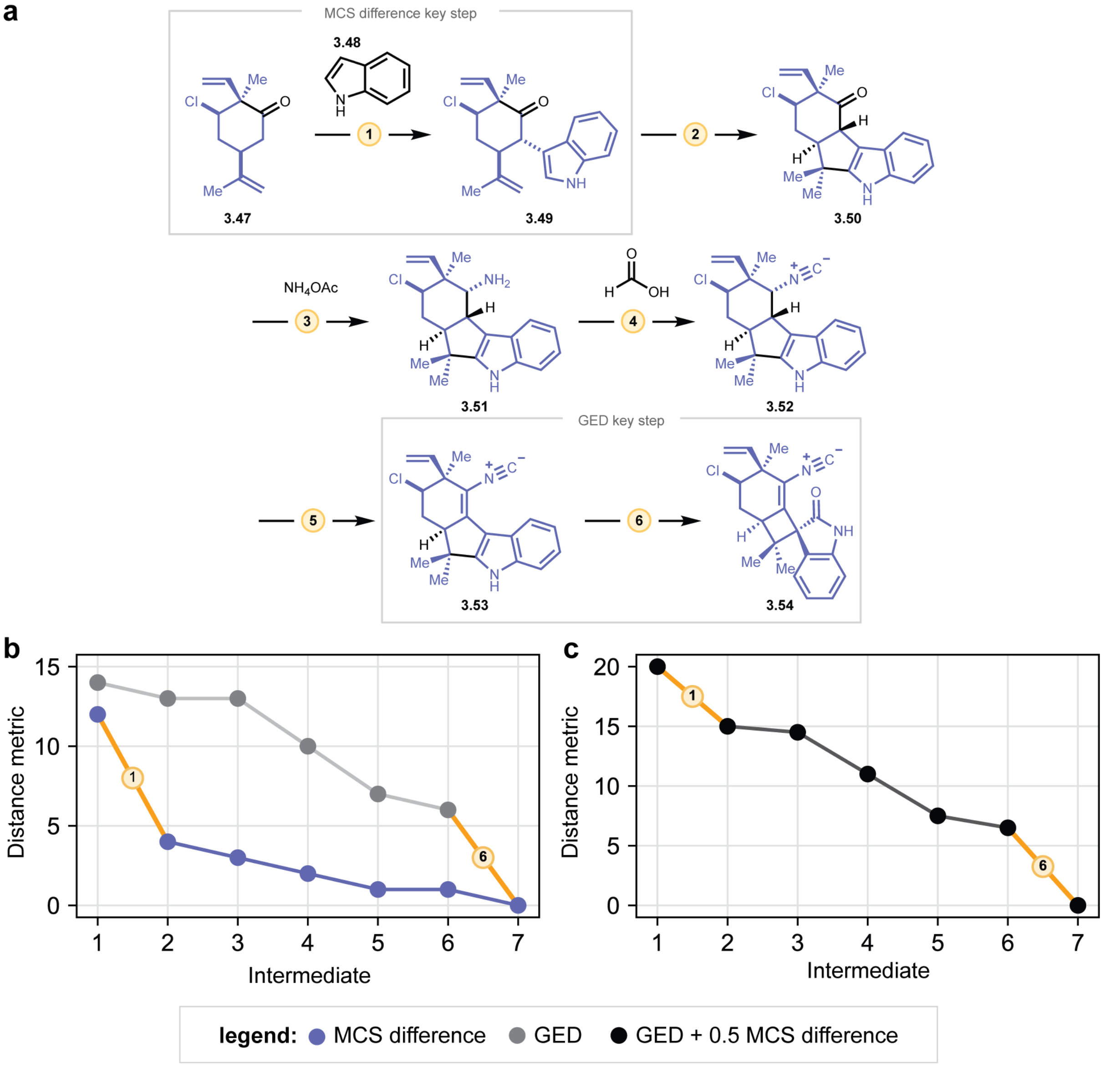


Figure 3.8. Incorporating MCS difference as an additional metric. **a.** Synthesis of welwitindolinone A (**3.54**) by Baran and Richter. Purple bonds indicate the maximum common substructure (MCS) between each intermediate and target. **b.** GED analysis (grey) assigns highest impact to step 6, while MCS analysis (purple) assigns this to step 1. **c.** A hybrid metric of GED + 0.5 MCS difference results in both important steps being identified.

3.5 Comparisons to other similarity metrics

For a wider perspective, we compared our synthetic route visualization method with other contemporary metrics, namely Tanimoto similarity, a popular measurement of fingerprint similarity[29–31], and Böttcher's complexity index C_m[32], as a representative of a complexity measurement that has seen precedence in recent graphing of total syntheses[13,33]. From Shenvi and coworkers' synthesis and brief literature review of himgaline (Fig. 3.9a, **3.55**)[13], we computed each intermediate's GED, Tanimoto similarity, and C_m similarity with the target, and inverted the Tanimoto and C_m scores to align their trends with our GED metric.

All three representations of the same route show different profiles (Fig. 3.9b) – both Tanimoto similarity and C_m labels assign highest impact to the last ketone reduction step (Fig. 3.9c), especially by a wide margin with Tanimoto similarity. In contrast, our hybrid GED and MCS difference metric highlights both the convergent step 3 (Fig. 3.9d) and pyridine ring reduction in step 5 (Fig. 3.9e). For another reference, the same analysis was performed on Larson and Sarpong's synthesis of **3.55**[34] (Fig. 3.9f). Once again, Tanimoto assigns the final carbonyl reduction as the highest-impact step (c.f. Fig. 3.9c), followed by deprotection step 15, which only breaks one concession bond, overlooking more significant steps such as cycloaddition step 2 (Fig. 3.9h), which was identified as highest impact when consulting both C_m and our difference metric, and reduction step 14 that removes three concession bonds (Fig. 3.9i).

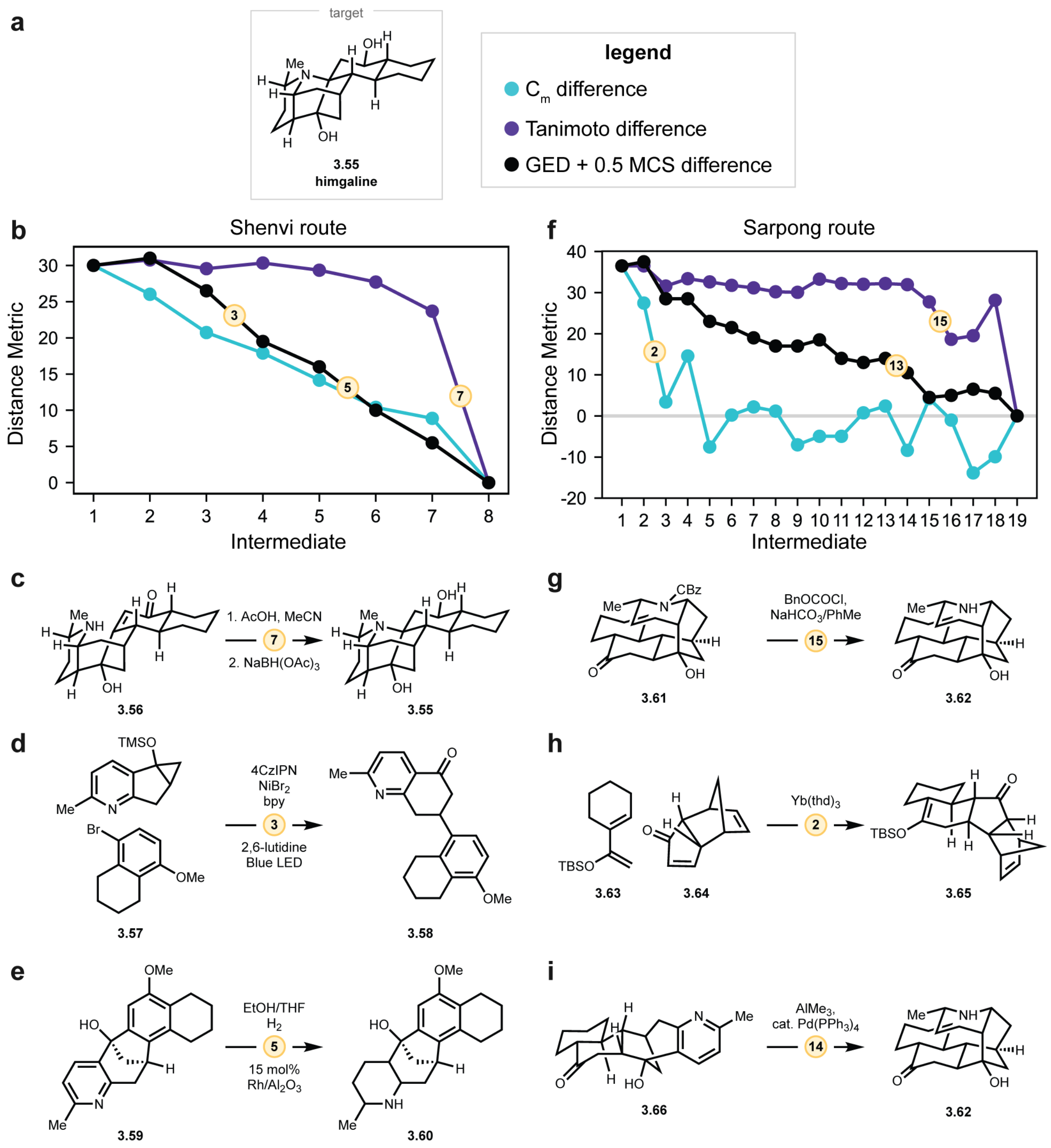


Figure 3.9. Comparision of various metrics for relative step impact evaluation. **a.** Structure of synthesis target himgaline (**3.55**). **b.** Overlaid plots of our distance metric, Tanimoto difference and C_m difference. All three metrics are normalized to have the same start and end points. **c.** Tanimoto similarity (as well as C_m) analysis evaluates the reduction step 7 as highest impact, drastically above the other steps. **d.** A convergent photoredox coupling highlighted by GED analysis. **e.** An aromatic ring reduction highlighted by our distance metric. **f.** Overlaid plots for Sarpong and coworkers' synthesis of **3.55. g.** Tanimoto analysis favors simple reduction steps over others that form more strategic bonds. **h.** A convergent cycloaddition that is highlighted by both C_m and our hybrid metric. **i.** A pyridine reduction, similar to **e**, highlighted by our distance metric. Structures of all intermediates in both synthetic routes can be found in Appendix C.

3.6 Conclusions and future direction

We applied matrix-encoded molecular systems and reactions developed in Chapter 1 towards analysis of multi-step reactions, and have developed a new visualization of total synthesis routes that can reveal the relative impacts of each individual steps. The matrix-based nature of this analysis method allows its conclusions to be visualized in a human-readable and understandable format. By analyzing many computer-proposed synthetic routes to the natural product stemoamide, we were able to extract high-impact key steps to incorporate into experiment, and discover two short asymmetric syntheses of this challenging target.

Moving forward from evaluating routes that are already published or generated, our method can also open avenues to discovery of novel reactions. Fig. 3.10a illustrates this vision with our published six-step synthesis of **3.18**, where several lower-impact functional group interconversions (Fig. 3.10a) can be bypassed with a strategic *anti*-Markovnikov hydroamidation that brings intermediate **3.35** directly to the target **3.18** (Fig. 3.10b).

A greater impact can be seen with Christmann's synthesis of englerin A[35] (Fig. 3.10c, **3.67**) – a protection and several low-impact steps are required to convert **3.68** to **3.72**, which undergoes ring-closing metathesis to form **3.73**, followed by deprotection to **3.74**. However, this can be greatly shortened via discovery of various bond-forming reactions, most effectively an alcohol-olefin metathesis that converts **3.68** directly to **3.74**.

Modern CASP has focused on reaction plausibility when proposing synthetic routes, leaning towards syntheses where all transformations have seen experimental precedent. Our method can complement this strategy by proposing novel reactions for discovery across points where a high-impact step can be made, or where several low-impact transformations can be conceivably omitted. We believe that improvements in modern computational data science and experimental methods will enable rapid discovery and optimization of reaction conditions, making room CASP programs to propose these valuable reactions.

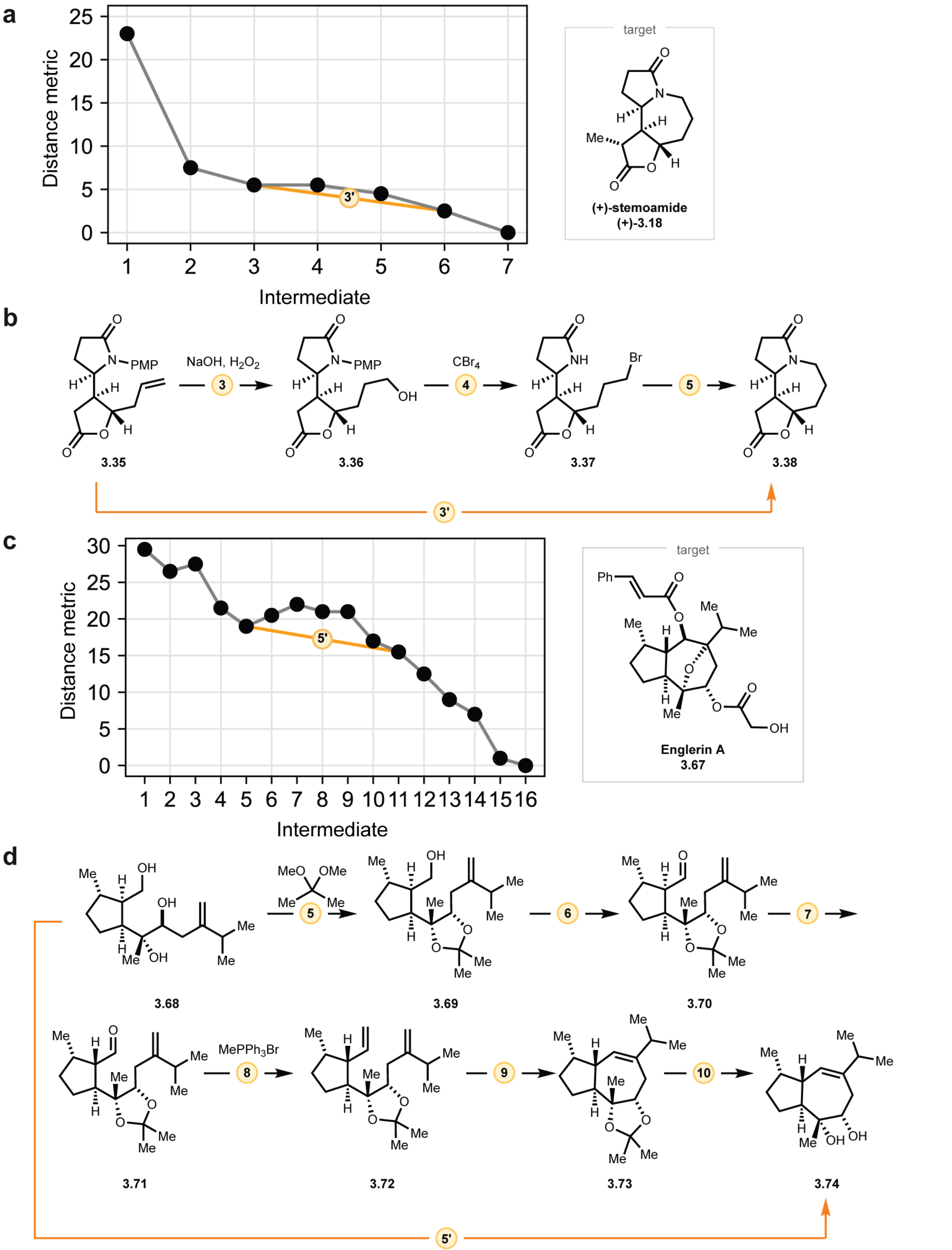


Figure 3.10. Plots of distance metrics enable proposals of novel reactions to lower step count. **a.** Visualization of our synthesis of **3.18**, where step 3', if invented, results in a four-step synthesis. **b.** An *anti*-Markovnikov hydroazidation enables bypass of three steps. **b.** Christmann and coworkers' synthesis of englerin A, where a hypothetical step 5' can reduce step count by five. **d.** An alcohol-olefin metathesis enables bypass of six steps. Structures of all intermediates in the synthesis of englerin A can be found in Appendix C.

3.7 References

(1) Nicolaou, K. C.; Edmonds, D. J.; Bulger, P. G. Cascade Reactions in Total Synthesis. *Angew. Chem. Int. Ed.* **2006**, *45* (43), 7134–7186. https://doi.org/10.1002/anie.200601872.

(2) Touré, B. B.; Hall, D. G. Natural Product Synthesis Using Multicomponent Reaction Strategies. *Chem. Rev.* **2009**, *109* (9), 4439–4486. https://doi.org/10.1021/cr800296p.

(3) Newhouse, T.; Baran, P. S.; Hoffmann, R. W. The Economies of Synthesis. *Chem Soc Rev* **2009**, *38* (11), 3010–3021. https://doi.org/10.1039/B821200G.

(4) Burns, N. Z.; Baran, P. S.; Hoffmann, R. W. Redox Economy in Organic Synthesis. *Angew. Chem. Int. Ed.* **2009**, *48* (16), 2854–2867. https://doi.org/10.1002/anie.200806086.

(5) Crossley, S. W. M.; Shenvi, R. A. A Longitudinal Study of Alkaloid Synthesis Reveals Functional Group Interconversions as Bad Actors. *Chem. Rev.* **2015**, *115* (17), 9465–9531. https://doi.org/10.1021/acs.chemrev.5b00154.

(6) Hendrickson, J. B. Systematic Synthesis Design. III. Scope of the Problem. *J. Am. Chem. Soc.* **1975**, *97* (20), 5763–5784. https://doi.org/10.1021/ja00853a022.

(7) Proudfoot, J. R. Reaction Schemes Visualized in Network Form: The Syntheses of Strychnine as an Example. *J. Chem. Inf. Model.* **2013**, *53* (5), 1035–1042. https://doi.org/10.1021/ci300556b.

(8) Schwan, J.; Christmann, M. Enabling Strategies for Step Efficient Syntheses. *Chem. Soc. Rev.* **2018**, *47* (21), 7985–7995. https://doi.org/10.1039/C8CS00399H.

(9) Barone, R.; Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 269–272. https://doi.org/10.1021/ci000145p.

(10) Whitlock, H. W. On the Structure of Total Synthesis of Complex Natural Products. *J. Org. Chem.* **1998**, *63* (22), 7982–7989. https://doi.org/10.1021/jo9814546.

(11) Scott, K. A.; Groch, J. R.; Bao, J.; Marshall, C. M.; Allen, R. A.; Nick, S. J.; Lauta, N. R.; Williams, R. E.; Qureshi, M. H.; Delost, M. D.; Njardarson, J. T. Minimalistic Graphical Presentation Approach for Total Syntheses. *Tetrahedron* **2022**, *126*, 133062. https://doi.org/10.1016/j.tet.2022.133062.

(12) Chanon, M.; Barone, R.; Baralotto, C.; Julliard, M.; Hendrickson, J. B. Information Theory Description of Synthetic Strategies in the Polyquinane Series. The Holosynthon Concept. *Synthesis* **1998**, *1998* (11), 1559–1583. https://doi.org/10.1055/s-1998-2191.

(13) Landwehr, E. M.; Baker, M. A.; Oguma, T.; Burdge, H. E.; Kawajiri, T.; Shenvi, R. A. Concise Syntheses of GB22, GB13, and Himgaline by Cross-Coupling and Complete Reduction. *Science* **2022**, *375* (6586), 1270–1274. https://doi.org/10.1126/science.abn8343.

(14) Heathcock, C. H.; Stafford, J. A. Daphniphyllum Alkaloids. 13. Asymmetric Total Synthesis of (-)-Secodaphniphylline. *J. Org. Chem.* **1992**, *57* (9), 2566–2574. https://doi.org/10.1021/jo00035a010.

(15) Jones, S. B.; Simmons, B.; Mastracchio, A.; MacMillan, D. W. C. Collective Synthesis of Natural Products by Means of Organocascade Catalysis. *Nature* **2011**, *475* (7355), 183–188. https://doi.org/10.1038/nature10232.

(16) Wang, L.; Wu, H.; Liu, C.; Jiang, T.; Yang, X.; Chen, X.; Tang, L.; Wang, Z. A Review of the Botany, Traditional Uses, Phytochemistry and Pharmacology of Stemonae Radix. *Phytochem. Rev.* **2022**, *21* (3), 835–862. https://doi.org/10.1007/s11101-021-09765-1.

(17) Yoritate, M.; Takahashi, Y.; Tajima, H.; Ogihara, C.; Yokoyama, T.; Soda, Y.; Oishi, T.; Sato, T.; Chida, N. Unified Total Synthesis of Stemoamide-Type Alkaloids by Chemoselective

Assembly of Five-Membered Building Blocks. *J. Am. Chem. Soc.* **2017**, *139* (50), 18386–18391. https://doi.org/10.1021/jacs.7b10944.

(18) Hou, Y.; Shi, T.; Yang, Y.; Fan, X.; Chen, J.; Cao, F.; Wang, Z. Asymmetric Total Syntheses and Biological Studies of Tuberostemoamide and Sessilifoliamide A. *Org. Lett.* **2019**, *21* (8), 2952–2956. https://doi.org/10.1021/acs.orglett.9b01042.

(19) Lin, Y.; Zhang, R.; Wang, D.; Cernak, T. Computer-Aided Key Step Generation in Alkaloid Total Synthesis. *Science* **2023**, *379* (6631), 453–457. https://doi.org/10.1126/science.ade8459.

(20) Hayashi, Y.; Tsuboi, W.; Ashimine, I.; Urushima, T.; Shoji, M.; Sakai, K. The Direct and Enantioselective, One-Pot, Three-Component, Cross-Mannich Reaction of Aldehydes. *Angew. Chem. Int. Ed.* **2003**, *42* (31), 3677–3680. https://doi.org/10.1002/anie.200351813.

(21) Notz, W.; Tanaka, F.; Watanabe, S.; Chowdari, N. S.; Turner, J. M.; Thayumanavan, R.; Barbas, C. F. The Direct Organocatalytic Asymmetric Mannich Reaction: Unmodified Aldehydes as Nucleophiles. *J. Org. Chem.* **2003**, *68* (25), 9624–9634. https://doi.org/10.1021/jo0347359.

(22) Brito, G. A.; Pirovani, R. V. Stemoamide: Total and Formal Synthesis. A Review. *Org. Prep. Proced. Int.* **2018**, *50* (3), 245–259. https://doi.org/10.1080/00304948.2018.1462032.

(23) Nicolaou, K. C.; Sarlah, D.; Shaw, D. M. Total Synthesis and Revised Structure of Biyouyanagin A. *Angew. Chem. Int. Ed.* **2007**, *46* (25), 4708–4711. https://doi.org/10.1002/anie.200701552.

(24) Gross, B. M.; Han, S.-J.; Virgil, S. C.; Stoltz, B. M. A Convergent Total Synthesis of (+)-Ineleganolide. *J. Am. Chem. Soc.* **2023**, *145* (14), 7763–7767. https://doi.org/10.1021/jacs.3c02142.

(25) Nicolaou, K. C.; Pan, S.; Shelke, Y.; Ye, Q.; Das, D.; Rigol, S. A Highly Convergent Total Synthesis of Norhalichondrin B. *J. Am. Chem. Soc.* **2021**, *143* (49), 20970–20979. https://doi.org/10.1021/jacs.1c10539.

(26) Gao, Y.; Ma, D. In Pursuit of Synthetic Efficiency: Convergent Approaches. *Acc. Chem. Res.* **2021**, *54* (3), 569–582. https://doi.org/10.1021/acs.accounts.0c00727.

(27) Urabe, D.; Asaba, T.; Inoue, M. Convergent Strategies in Total Syntheses of Complex Terpenoids. *Chem. Rev.* **2015**, *115* (17), 9207–9231. https://doi.org/10.1021/cr500716f.

(28) Baran, P. S.; Richter, J. M. Enantioselective Total Syntheses of Welwitindolinone A and Fischerindoles I and G. *J. Am. Chem. Soc.* **2005**, *127* (44), 15394–15396. https://doi.org/10.1021/ja056171r.

(29) Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction*; International Business Machines Corporation, 1958.

(30) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminformatics* **2015**, *7* (1), 20. https://doi.org/10.1186/s13321-015-0069-3.

(31) Safizadeh, H.; Simpkins, S. W.; Nelson, J.; Li, S. C.; Piotrowski, J. S.; Yoshimura, M.; Yashiroda, Y.; Hirano, H.; Osada, H.; Yoshida, M.; Boone, C.; Myers, C. L. Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical–Genetic Interactions. *J. Chem. Inf. Model.* **2021**, *61* (9), 4156–4172. https://doi.org/10.1021/acs.jcim.0c00993.

(32) Böttcher, T. An Additive Definition of Molecular Complexity. *J. Chem. Inf. Model.* **2016**, *56* (3), 462–470. https://doi.org/10.1021/acs.jcim.5b00723.

(33) Demoret, R. M.; Baker, M. A.; Ohtawa, M.; Chen, S.; Lam, C. C.; Khom, S.; Roberto, M.; Forli, S.; Houk, K. N.; Shenvi, R. A. Synthetic, Mechanistic, and Biological Interrogation of

Ginkgo Biloba Chemical Space En Route to (−)-Bilobalide. *J. Am. Chem. Soc.* **2020**, *142* (43), 18599–18618. https://doi.org/10.1021/jacs.0c08231.

(34) Larson, K. K.; Sarpong, R. Total Synthesis of Alkaloid (±)-G. B. 13 Using a Rh(I)-Catalyzed Ketone Hydroarylation and Late-Stage Pyridine Reduction. *J. Am. Chem. Soc.* **2009**, *131* (37), 13244–13245. https://doi.org/10.1021/ja9063487.

(35) Willot, M.; Radtke, L.; Könning, D.; Fröhlich, R.; Gessner, V. H.; Strohmann, C.; Christmann, M. Total Synthesis and Absolute Configuration of the Guaiane Sesquiterpene Englerin A. *Angew. Chem. Int. Ed.* **2009**, *48* (48), 9105–9108. https://doi.org/10.1002/anie.200905032.

Chapter 4 Development of an Automated Platform for High-Throughput Experimentation

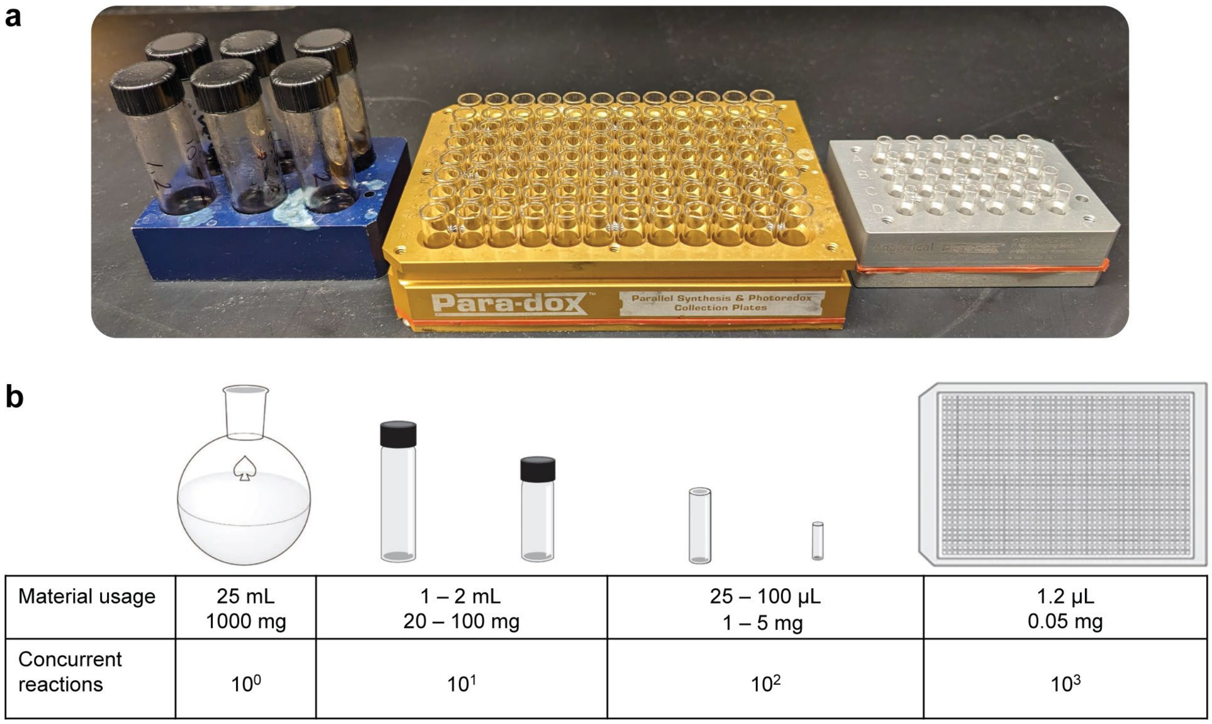
4.1 Introduction to High-Throughput Experimentation

Chapters 2 and 3 showed how matrix techniques can be applied in tandem with other computational methods to search for valuable reactions to discover. This chapter will discuss how our group turns these computer recommendations into experimental methods.

4.1.1 Background

Our group places high emphasis on high-throughput experimentation (HTE), the performance of many experiments in parallel, frequently in a wellplate format (Fig. 4.1a). The adaptation of HTE in chemistry research is becoming increasingly widespread, ranging from drug development[1–5] and reaction discovery[6–10] to the preparation of large-scale datasets[11,12].

Compared to traditional benchtop synthesis, HTE is capable of screening a wider range of substrates and experimental conditions per unit reagent used, while being more economical and environmentally friendly in terms of chemical requirement[13]. While synthesis in flasks and 1-2 dram vials can be parallelized to some extent, they nonetheless require milligram to gram-scale amounts of reagent to generate fewer data points (Fig. 4.1b). With the same physical footprint, shell and microvials arrayed in wellplates require much less material per well, while performing a hundred to over a thousand reactions. This generation of organized data in high volume couples well with the growing adaptation of data science methods to enable rapid discoveries of new reactivities[14–17].



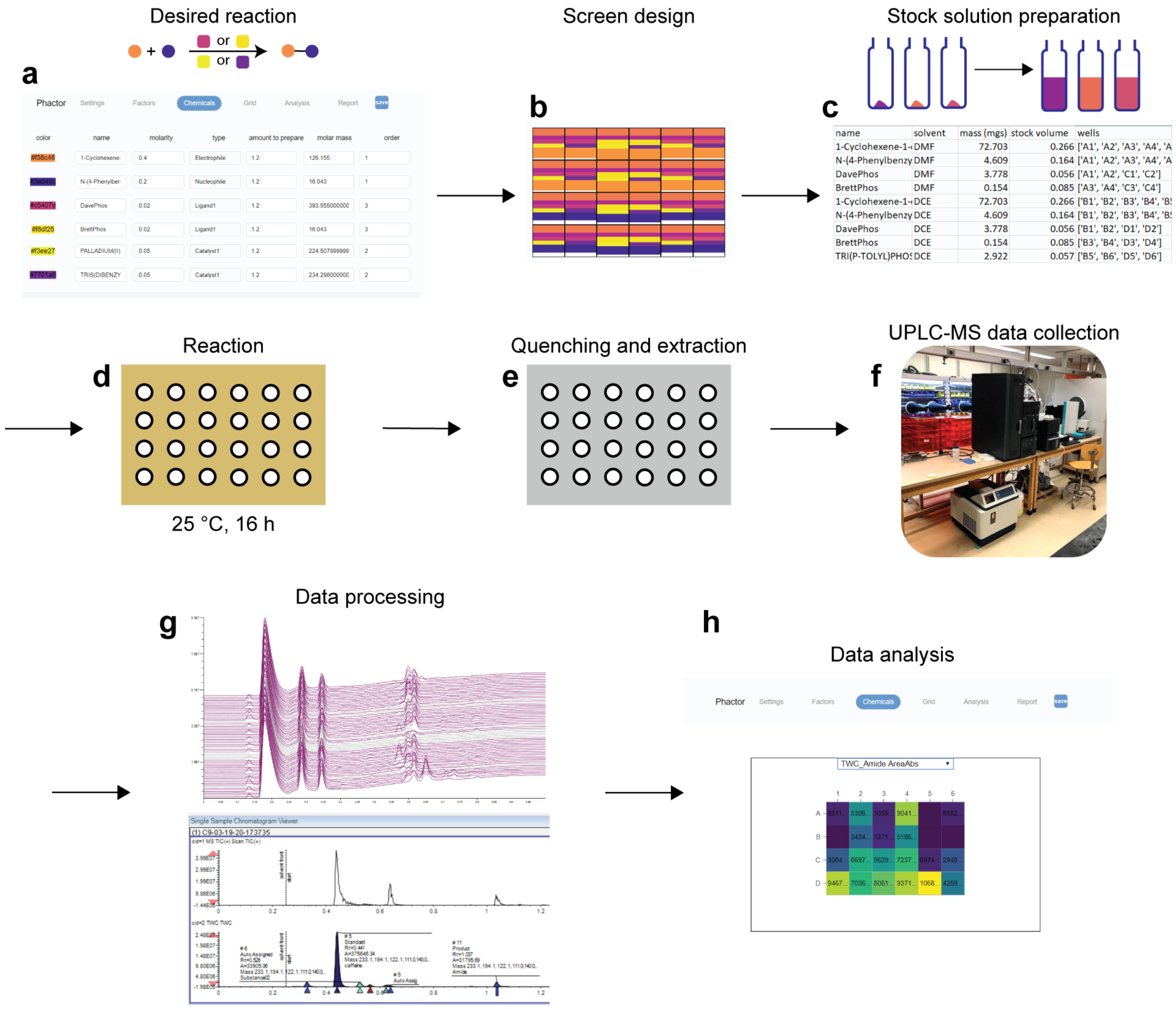
Material usage	25 mL 1000 mg	1 – 2 mL 20 – 100 mg	25 – 100 µL 1 – 5 mg	1.2 µL 0.05 mg
Concurrent reactions	10^0	10^1	10^2	10^3

Figure 4.1. Labware used for HTE. **a.** Examples of labware used to execute reactions in parallel. From left to right: capped glass vials for 0.5–2 mL scale, glass shell vials for 100–300 µL scale, and micro vials for 10–30 µL scale. **b.** Comparison of material usage and concurrent reactions permitted for each type of labware.

4.1.2 Overview of current HTE workflow

To support our HTE screening infrastructure, we developed phactor™[18], an online software that couples with our digital chemical inventory to support design, execution, analysis and data storage of HTE screens. The workflow of a typical HTE screen is outlined in Fig. 4.2.

First, the user, with a desired reaction in mind, selects the types and amounts of reagents to screen (Fig. 4.2a). With this information, a screen design can be constructed manually, or automatically populated by phactor™(4.2b), such that all combinations of reaction conditions are represented in the grid. Next, a downloadable recipe will be produced, containing directions such as preparations of all stock solutions, the wells in which they should be dosed, and the dosage volume (4.2c). The prepared screen is then moved to a suitable reactor to be heated and stirred at the prescribed conditions (4.2d). After the reaction is complete, it is quenched and extracted (4.2e), and then an analytical plate prepared and analyzed with UPLC-MS (4.2f). An external software, Virscidian, enables characterization of LC-MS traces and export of results (4.2g), which are imported back into phactor™ for visualization and guidance on future experiment directions (4.2h).



color	name	molarity	type	amount to prepare	molar mass	order
#f58c46	1-Cyclohexene-	0.4	Electrophile	1.2	126.155	1
#3e049c	N-(4-Phenylben	0.2	Nucleophile	1.2	16.043	1
#c5407e	DavePhos	0.02	Ligand1	1.2	393.555000000	3
#f8df25	BrettPhos	0.02	Ligand1	1.2	16.043	3
#f3ee27	PALLADIUM(II)	0.05	Catalyst1	1.2	224.507999999	2
#7701a8	TRIS(DIBENZY	0.05	Catalyst1	1.2	234.298000000	2

name	solvent	mass (mgs)	stock volume	wells
1-Cyclohexene-1-(	DMF	72.703	0.266	['A1', 'A2', 'A3', 'A4', 'A
N-(4-Phenylbenzy	DMF	4.609	0.164	['A1', 'A2', 'A3', 'A4', 'A
DavePhos	DMF	3.778	0.056	['A1', 'A2', 'C1', 'C2']
BrettPhos	DMF	0.154	0.085	['A3', 'A4', 'C3', 'C4']
1-Cyclohexene-1-(	DCE	72.703	0.266	['B1', 'B2', 'B3', 'B4', 'B!
N-(4-Phenylbenzy	DCE	4.609	0.164	['B1', 'B2', 'B3', 'B4', 'B!
DavePhos	DCE	3.778	0.056	['B1', 'B2', 'D1', 'D2']
BrettPhos	DCE	0.154	0.085	['B3', 'B4', 'D3', 'D4']
TRI(P-TOLYL)PHOS	DCE	2.922	0.057	['B5', 'B6', 'D5', 'D6']

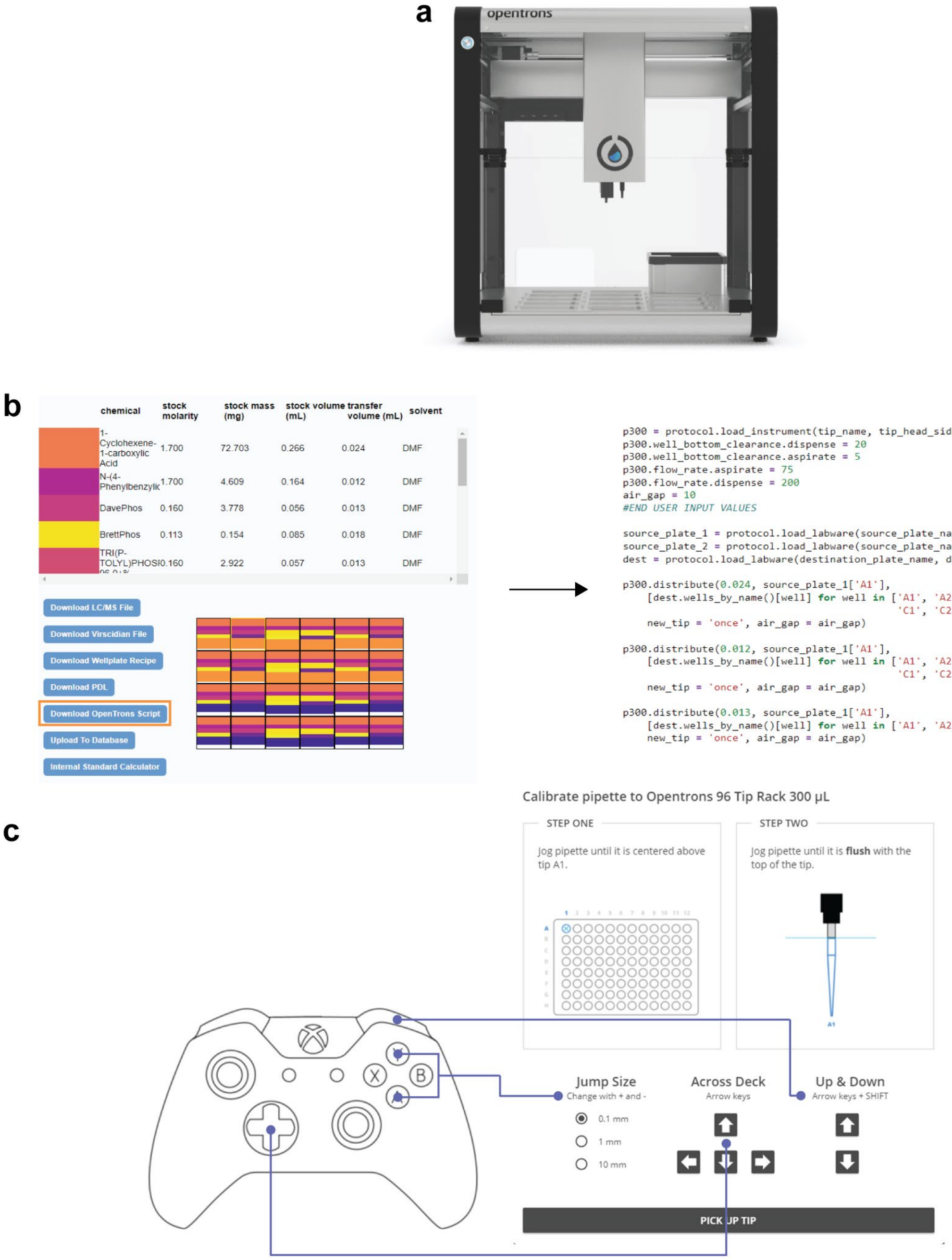
	1	2	3	4	5	6
A	8511...	5306...	1059...	9041...		8552...
B		3434...	1271...	5186...		
C	3084...	6697...	5629...	7237...	8974...	2948...
D	9467...	7036...	8061...	9371...	1068...	4269...

Figure 4.2. Workflow of an HTE screen. **a.** phactor™ interface for reagent selection. Users enter experimental data such as reagent identity, reagent class, desired final concentration, and overage. **b.** Reagent grid, which users can manually or automatically generate. Each colored bar represents on reagent present in the indicated well. **c.** Screen recipe spreadsheet generated by phactor™ for stock solution preparation. **d.** After dosing, the wellplate is moved to an appropriate environment for stirring and/or heating **e.** Wellplate is quenched and transferred to an analytical plate after reaction time has passed. **f.** High-throughput analysis is performed by UPLC-MS. **g.** Raw LC-MS data being analyzed with third-party software **h.** Analysis results can be uploaded into phactor™ for visualization.

4.2 Expanding the extent of automation through integration of liquid handler

Our group has employed phactor™ to conduct a wide range of HTE screens, discovering many novel reactivities and conducting direct-to-biology campaigns. To further bolster the efficiency of this software, we sought to introduce automation at the screen preparation and workup stage through integrating phactor™'s screen designs with the Opentrons OT-2 (Fig. 4.3a), a robotic liquid handler capable of transferring reagents as liquids, solutions or suspensions across many types of wellplates. The OT-2 was selected due to its lower cost and smaller footprint compared to its counterparts, while being equipped with a Python API, which allows operations to be executed via procedurally generated scripts instead of needing the user to manually operate a separate interface.

We have implemented a new functionality into phactor™ which allows a designed screen to be exported into a Python protocol compatible with the liquid handler. This protocol only requires minor user input on the size and location of wellplates, after which it can be directly imported into the Opentrons software and executed by the robot (Fig. 4.3b). To facilitate pre-run calibration in setups where the computer is not directly beside the robot, a video game controller can be used to interface with directional controls on the Opentrons software (Fig. 4.3 c).



	chemical	stock molarity	stock mass (mg)	stock volume (mL)	transfer volume (mL)	solvent
🟧	1-Cyclohexene-1-carboxylic Acid	1.700	72.703	0.266	0.024	DMF
🟪	N-(4-Phenylbenzylic	1.700	4.609	0.164	0.012	DMF
🟥	DavePhos	0.160	3.778	0.056	0.013	DMF
🟨	BrettPhos	0.113	0.154	0.085	0.018	DMF
🟥	TRI(P-TOLYL)PHOS[0.160		2.922	0.057	0.013	DMF

Download LC/MS File
Download Virscidian File
Download Wellplate Recipe
Download PDL
Download OpenTrons Script
Upload To Database
Internal Standard Calculator

```
p300 = protocol.load_instrument(tip_name, tip_head_side, tip_racks=[tiprack_1])
p300.well_bottom_clearance.dispense = 20
p300.well_bottom_clearance.aspirate = 5
p300.flow_rate.aspirate = 75
p300.flow_rate.dispense = 200
air_gap = 10
#END USER INPUT VALUES

source_plate_1 = protocol.load_labware(source_plate_name, source_plate_location_1)
source_plate_2 = protocol.load_labware(source_plate_name, source_plate_location_2)
dest = protocol.load_labware(destination_plate_name, destination_plate_location)

p300.distribute(0.024, source_plate_1['A1'],
    [dest.wells_by_name()[well] for well in ['A1', 'A2', 'A3', 'A4', 'A5', 'A6',
                                             'C1', 'C2', 'C3', 'C4', 'C5', 'C6']],
    new_tip = 'once', air_gap = air_gap)

p300.distribute(0.012, source_plate_1['A1'],
    [dest.wells_by_name()[well] for well in ['A1', 'A2', 'A3', 'A4', 'A5', 'A6',
                                             'C1', 'C2', 'C3', 'C4', 'C5', 'C6']],
    new_tip = 'once', air_gap = air_gap)

p300.distribute(0.013, source_plate_1['A1'],
    [dest.wells_by_name()[well] for well in ['A1', 'A2', 'C1', 'C2']],
    new_tip = 'once', air_gap = air_gap)
```

Calibrate pipette to Opentrons 96 Tip Rack 300 µL

STEP ONE

Jog pipette until it is centered above tip A1.

STEP TWO

Jog pipette until it is **flush** with the top of the tip.

Jump Size
Change with + and -

● 0.1 mm
○ 1 mm
○ 10 mm

Across Deck
Arrow keys

Up & Down
Arrow keys + SHIFT

Figure 4.3. Implementing automated screen export and execution. **a.** Photograph of Opentrons OT-2 autopipettor robot. **b.** New feature added to phactor™ indicated by orange rectangle, allowing export of screen design into OT-2 compatible python scripts. **c.** Input mappings for OT-2 robot calibration using a video game controller[19].

4.3 Assessing performance of robot compared to manual preparation of HTE screens

4.3.1 Homogeneous reagent solutions

To compare the outcome of an HTE screen executed by a human with that of a robot, a 96-wellplate was split into two halves – one with all reagents dosed by hand, the other with the robot. A simple amide coupling reaction was chosen as a robust reaction where all reagents are soluble in N,N'-dimethylformamide (DMF), a high-boiling solvent suitable for HTE screens. To evaluate the reproducibility of results, a total of 8 conditions are tested over 48 wells, hence each condition is replicated 6 times.

A heatmap of product yield, measured by concentration relative to an internal standard, is illustrated in Fig. 4.4a, accompanied by a bar chart in 4.4b comparing, for each condition, the product yielded through manual verses automatic dosing. In conditions that produce higher yield, manual dosing showed better performance, while lower-yielding conditions showed approximately equal results. Notably, even in conditions with little observed yield, both manual and robot-dosed wells recorded non-zero yields. This is an important observation, as HTE is frequently used in reaction discovery, where initial results may occur with very low yields as they are yet to be optimized. Being capable of detecting these products means that these initial successes are not ruled out due to false negatives.

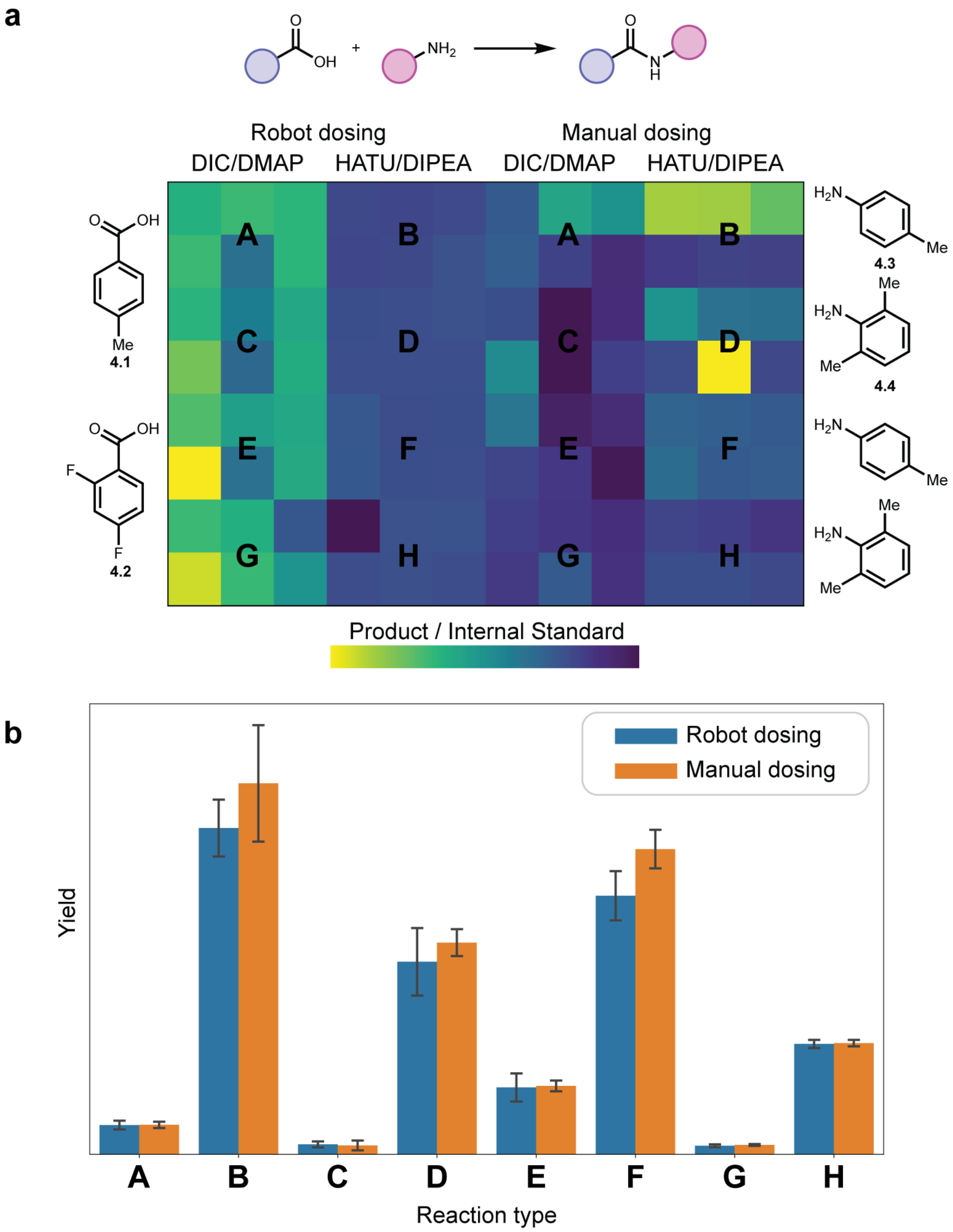


Figure 4.4. Accessing autopipetter robot performance with a homogeneous reaction**. a.** Reagents and conditions used. Each unique condition is replicated six times in a 2 x 3 grid. Heatmap shows analytical yield relative to internal standard (1 mg/mL caffeine). **b.** Comparison of yields between wells dosed automatically and manually. DIC = N,N'-diisopropylcarbodiimide, DMAP = 4-dimethylaminopyridine, HATU = hexafluorophosphate azabenzotriazole tetramethyl uranium, DIPEA = N,N'-diisopropylethylamine.

4.3.2 Heterogeneous reagents

In an HTE campaign, all reagent stock solutions are prepared in concentrations that are several times higher than the desired concentration in the reaction flask, since they are diluted when dosed into a well with other stock solutions. Reagents that are not sufficiently soluble in the reaction solvent will hence form slurries or suspensions, which may not be dosed accurately. Vigorous agitation of the stock solutions is used to alleviate this issue in both manual and automated workflows, but we wished to explore an alternative method of dosing selected reagents in solid form. Single-use polypropylene scoops have been used for this purpose, but their sizes are limited, so reagents may not be dosed accurately, or require multiple scoops.

A Buchwald-Hartwig coupling screen[20] was selected for the presence of two poorly soluble bases, sodium tert-butoxide and cesium carbonate. Using computer-aided design (CAD) software, we modelled a simple hemispheric scoop with an editable diameter (Fig. 4.5a). Scoops with a range of diameters are then 3D printed (Fig. 4.5b) and the weight of a single flat scoop of the desired reagent is recorded. Four reagent weights are recorded per scoop size, with the first measurement discarded as the 3D printing process produces small grooves in between material layers that traps chemical powder. Plotting measured masses against scoop volume and performing linear best fit yields a calibration line (Fig. 4.5c, blue dots), which is used to calculate the desired scoop diameter. To verify performance of the resultant printed scoops, a final round of weighing was performed, and both solids were able to be dosed in the appropriate amount (Fig. 4.5c, orange dots).

Following the methodology in Chapter 4.3.1, eight reaction conditions were selected and another 96-well screen was run, but with only the bases manually dosed on the right half of the wellplate. All other reagents, including bases for the left half of the wellplate, were dosed by the OT-2 robot as stock solutions which were continually agitated with a magnetic stirrer on the deck. Reaction yields are visualized as a heatmap in Fig. 4.6a, and bar charts in 4.6b. Unlike the homogeneous reaction, wells with sodium tert-butoxide performed better when the base was dosed automatically, while wells with cesium carbonate performed better with manual dosing. We believe the cause for this difference lies in the different particle sizes of these bases, as well as their behavior in the solvent, toluene. Sodium tert-butoxide forms a uniform suspension in toluene, and can hence be reliably dosed with a pipette while the solution is being agitated. However, as it is a fine powder, the manual dosing process is susceptible to reagent loss caused by static electricity, which is present under the glovebox environment in which this reaction was performed. The opposite is true for cesium carbonate, which forms a hard cake in toluene, rendering the stock solution difficult to agitate. Any aspirated liquid will hence have lower concentration of base than expected. Since cesium carbonate is a coarser powder, it is less affected by static electricity, leading to more accurate dosing in solid form.

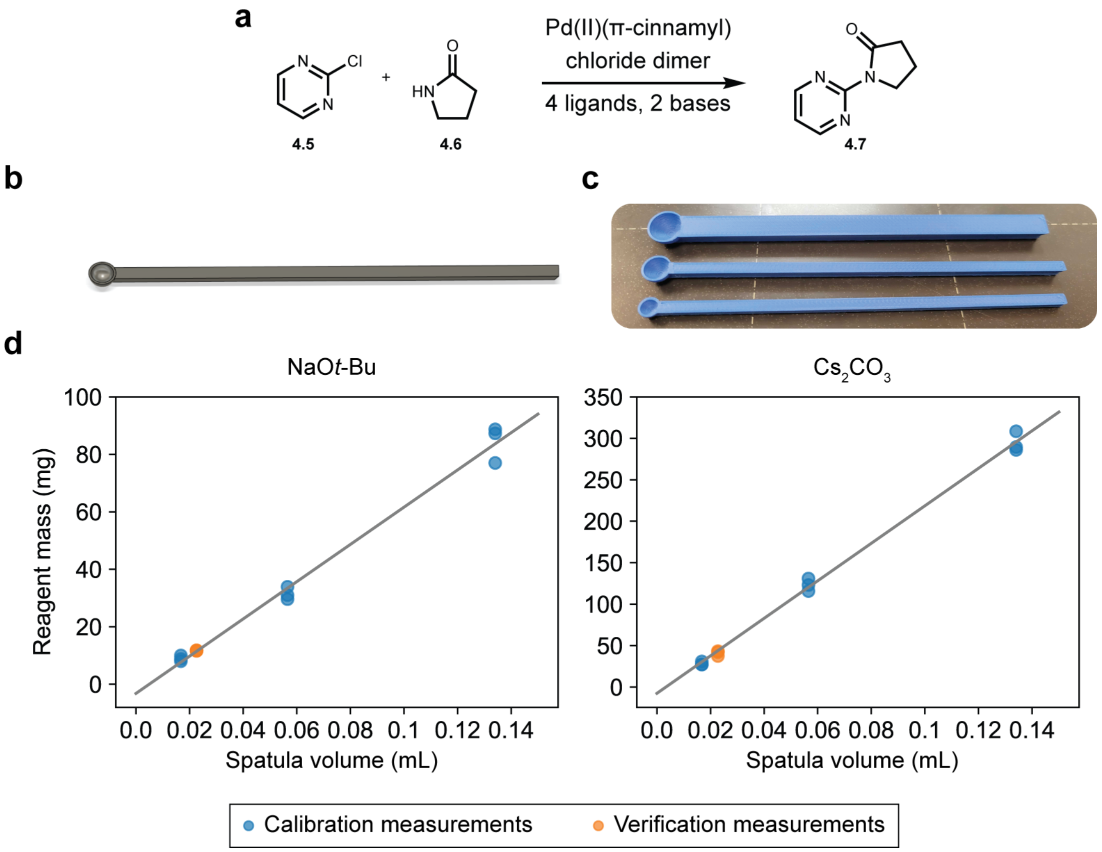


Figure 4.5. Exploration of manual solid dosing when reagents are poorly soluble. **a.** A Buchwald-Hartwig coupling screen adapted from a publication from Wood and coworkers. **b.** CAD model of a scoop with a hemispherical head. **c.** 3D printed scoops with various radii. **d.** Process to determine optimal scoop size. Scoops with radii 2mm, 3mm and 4mm are used for calibration. A plot of weighed reagent mass against volume produces a calibration curve, from which the ideal volume and subsequently radius can be calculated. A scoop printed with the desired radius is used to weigh reagent again, for performance verification.

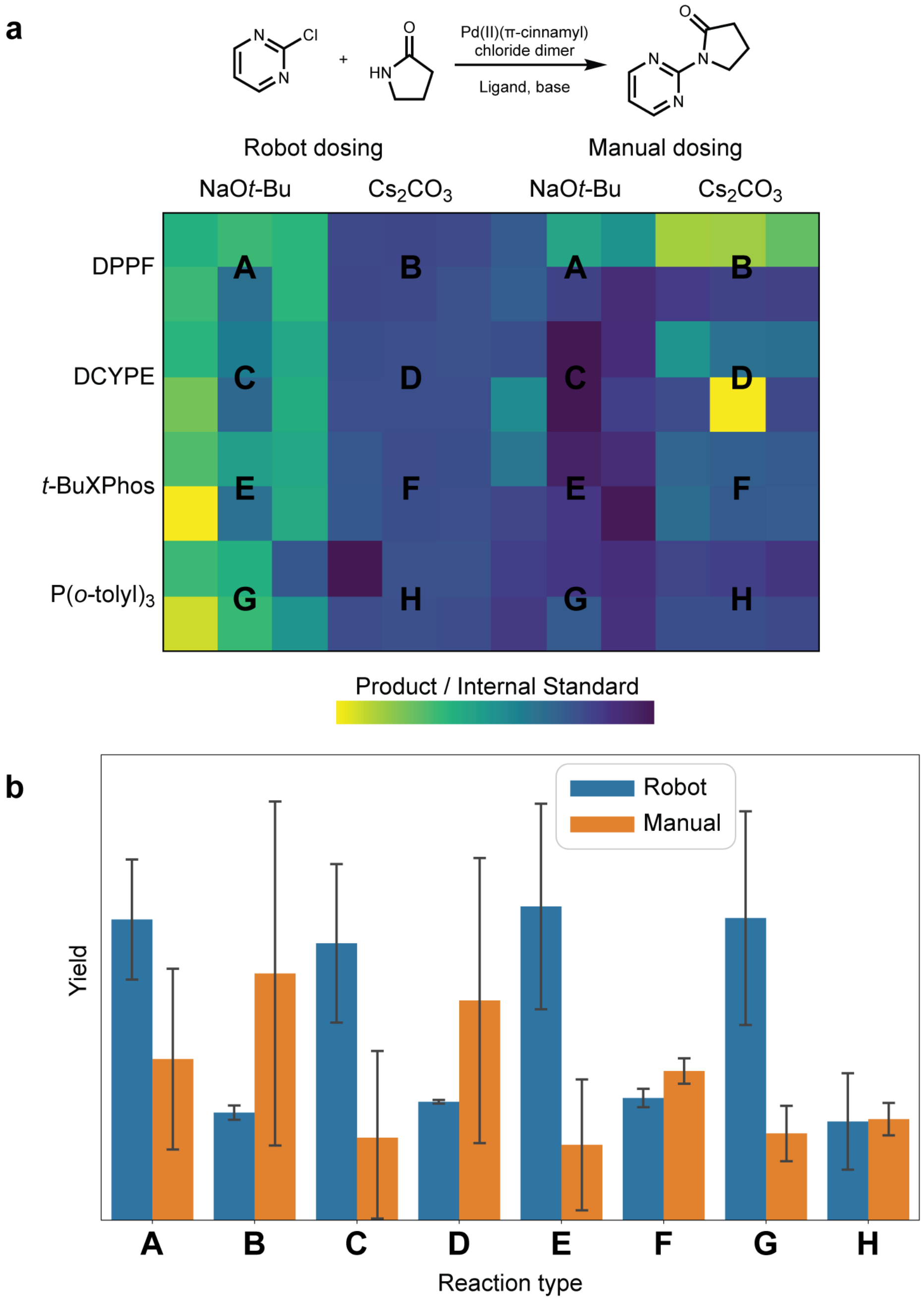


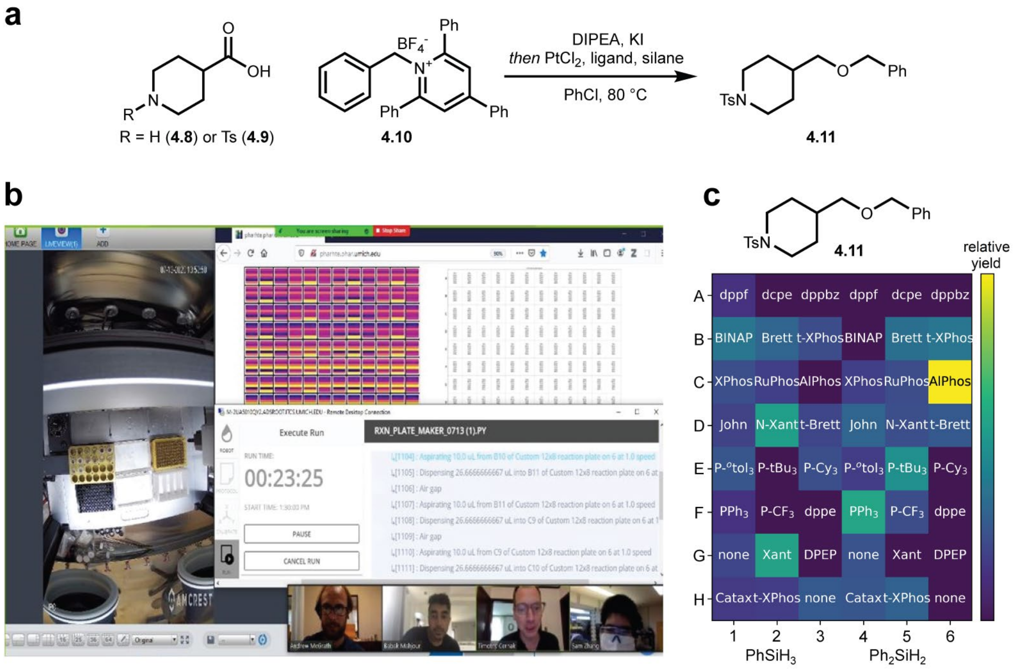
Figure 4.6. Accessing autopipetter robot performance using a reaction with two poorly soluble bases. **a.** Reagents and conditions used. Each unique condition is replicated six times in a 2 x 3 grid. Heatmap shows analytical yield relative to internal standard (1 mg/mL caffeine). **b.** Comparison of yields between wells with the base dosed automatically and manually. DPPF = 1,1′ -Ferrocenediyl-bis(diphenylphosphine), DCYPE = 1,2-Bis(dicyclohexylphosphino)ethane.

4.4 Applications of online HTE platform

4.4.1 Remote experiment collaboration over teleconferencing software

During the COVID-19 pandemic throughout 2020, laboratory occupancy was heavily restricted to reduce spread of the disease. This highlighted the potential for automated experimentation systems[21], and provided an opportunity to showcase the potential for remote collaboration enabled by our HTE infrastructure.

One of our projects disrupted by laboratory lockdown was the optimization of a newly discovered amine–acid etherification reaction (Fig. 4.7a). After observation of the initial hit, a reaction condition screen was designed online in phactor™ for 2 substrates, 24 ligands and 2 reductants. This recipe is then accessed and prepared in the lab building by another member. A camera mounted above the OT-2 deck and connected to the controlling computer allowed other project participants to observe the experiment and provide feedback in real time through teleconferencing software (Fig. 4.7b). Analysis of this screen revealed AlPhos and diphenylsilane as the best combination of ligand and reductant, and that tosyl protection of the piperidine ring on the acid substrate (**4.9**) was necessary, as no product was observed with the free amine.



M-2UA5010QV2.ADSROOT.ITCS.UMICH.EDU - Remote Desktop Connection

RXN_PLATE_MAKER_0713 (1).PY

```
↳[1104] : Aspirating 10.0 uL from B10 of Custom 12x8 reaction plate on 6 at 1.0 speed
↳[1105] : Dispensing 26.6666666667 uL into B11 of Custom 12x8 reaction plate on 6 at
↳[1106] : Air gap
↳[1107] : Aspirating 10.0 uL from B11 of Custom 12x8 reaction plate on 6 at 1.0 speed
↳[1108] : Dispensing 26.6666666667 uL into C9 of Custom 12x8 reaction plate on 6 at 1
↳[1109] : Air gap
↳[1110] : Aspirating 10.0 uL from C9 of Custom 12x8 reaction plate on 6 at 1.0 speed
↳[1111] : Dispensing 26.6666666667 uL into C10 of Custom 12x8 reaction plate on 6 at
```

	1	2	3	4	5	6
A	dppf	dcpe	dppbz	dppf	dcpe	dppbz
B	BINAP	Brett	t-XPhos	BINAP	Brett	t-XPhos
C	XPhos	RuPhos	AlPhos	XPhos	RuPhos	AlPhos
D	John	N-Xant	t-Brett	John	N-Xant	t-Brett
E	P-^{o}tol$_3$	P-tBu$_3$	P-Cy$_3$	P-^{o}tol$_3$	P-tBu$_3$	P-Cy$_3$
F	PPh$_3$	P-CF$_3$	dppe	PPh$_3$	P-CF$_3$	dppe
G	none	Xant	DPEP	none	Xant	DPEP
H	Catax	t-XPhos	none	Catax	t-XPhos	none
	$PhSiH_3$			Ph_2SiH_2		

Figure 4.7. Remote collaboration enabled by integrating phactor™'s online screen design storage with automated screen execution with an Opentrons autopipetter. **a.** Schematic of amine–acid ether synthesis. **b.** Teleconferencing software enables real-time experiment monitoring over teleconferencing software. **c.** Product yield of screen across 24 ligands and 2 reductants. Only half of the plate is shown, as all wells with unprotected acid substrate (**4.8**) did not yield any product.

4.4.2 Small-scale library synthesis

The flexibility of phactor™'s screen design capability enables the execution of novel HTE screens. We have discovered a deaminative esterification reaction between an *sp*^{3} amine activated as the triphenylpyridinium (TPP) salt and a carboxylic acid, and sought to probe its utility in late-stage functionalization by performing a small-scale library synthesis. Amlodipine was selected as the activated amine drug (**4.12**) to be coupled with 96 diverse acids. All reagents except the acids were weighed into 8 mL vials, while the 96 acids were weighed into glass shell vials arrayed in a wellplate. All reagents were brought into a nitrogen-filled glovebox for stock solution preparation. The robotic platform confers several advantages in this setup – both the activated amine and several acid substrates were not fully soluble at its stock solution concentration, but an on-deck tumble stirrer can be employed to agitate the amine, while mixing cycles can be programmed before the aspiration of each acid to induce suspension of any insoluble particles. In addition, as these acid solutions were in a grid, they can be rapidly dosed into the reaction wellplate using a multi-channel pipette (Fig. 4.8b).

The preparation of stock solutions was also greatly accelerated. As it is difficult to manually weigh out an exact amount of solid, users can input their measured weight for each reagent's into phactor™, which will compute the revised volume of solvent required for all stock solutions to be their desired concentration. Manually adjusting a micropipette for over 100 reagents would have been time consuming for the large amount of reagents in this screen, and introduces an additional source of error. Instead, the volumes required are imported into a custom OT-2 script, which directs the autopipetter to dose the appropriate amount of solvent into each well to reach the desired 0.30 M concentration (Fig. 4.8c). Analysis of the screen results revealed that majority of wells yielded the desired ester product (Fig. 4.8d), and benchtop scale-up of selected wells produced satisfactory yields.

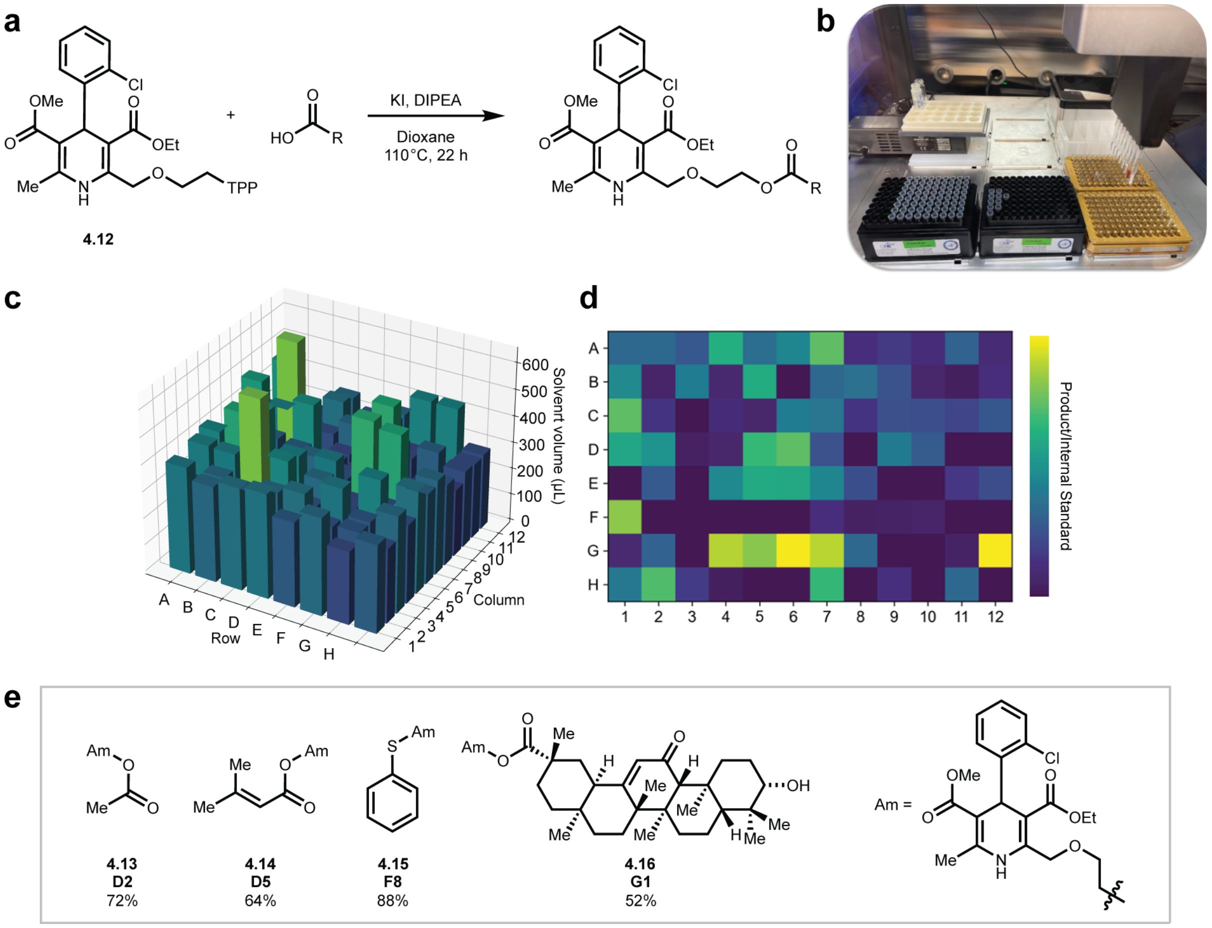


Figure 4.8. Library synthesis with an amine–acid esterification reaction. **a.** Reaction scheme. **b.** Screen setup using the Opentrons OT-2. The multipipette head can quickly transfer all 96 acids from the wellplate in which they were prepared to the one dosed with all other reagents. **c.** Volume of solvent added into each of the 96 shell vials containing various acid substrates. Automated liquid handling enabled efficient dosing of these varied volumes. **d.** Heatmap of product yield, measured by UV peak integration against internal standard. As products have varied UV absorbance, the heatmap does not represent exact relative yield. Distribution of acids are as follows. Rows A and B: (hetero)aryl acids, C: acetic acids, D: aliphatic acids, E: Protected amino acids, F: non-carboxylic acids, G: carboxylic acid-containing drugs, H: carboxylate salts. Identities of all 96 acids are displayed in Appendix C.

4.5 Conclusions and future work

We have built upon our online HTE planning software phactor™ to allow exporting of experimental screens to an Opentrons OT-2 liquid handler, which further extends our capability to automate execution of HTE screens. Side-by-side comparisons with manual screen setup reveal that both methods perform similarly in terms of product yield, affirming that the move towards automated HTE screen setup will not compromise on the quality of data produced, while reducing the random error inherent to manual pipetting into dense wellplates. An exception is made with solids that do not form a uniform suspension upon agitation, which may warrant manual dosing with custom-made spatula.

We are currently expanding our automation scope to include ultraHTE screens, which are conducted under nanoscale in 384- or 1536-wellplates. The OT-2 has already demonstrated reliability in preparation of stock solution in 384-wellplates, which are moved onto other instruments for precision dosing and heating. Robotic devices for precise positioning of labware across several instruments have potential for implementation in this space, in order to further increase throughput of these instruments.

4.6 References

(1) Mennen, S. M.; Alhambra, C.; Allen, C. L.; Barberis, M.; Berritt, S.; Brandt, T. A.; Campbell, A. D.; Castañón, J.; Cherney, A. H.; Christensen, M.; Damon, D. B.; Eugenio De Diego, J.; García-Cerrada, S.; García-Losada, P.; Haro, R.; Janey, J.; Leitch, D. C.; Li, L.; Liu, F.; Lobben, P. C.; MacMillan, D. W. C.; Magano, J.; McInturff, E.; Monfette, S.; Post, R. J.; Schultz, D.; Sitter, B. J.; Stevens, J. M.; Strambeanu, I. I.; Twilton, J.; Wang, K.; Zajac, M. A. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Org. Process Res. Dev.* **2019**, *23* (6), 1213–1242. https://doi.org/10.1021/acs.oprd.9b00140.

(2) Cernak, T.; Gesmundo, N. J.; Dykstra, K.; Yu, Y.; Wu, Z.; Shi, Z.-C.; Vachal, P.; Sperbeck, D.; He, S.; Murphy, B. A.; Sonatore, L.; Williams, S.; Madeira, M.; Verras, A.; Reiter, M.; Lee, C. H.; Cuff, J.; Sherer, E. C.; Kuethe, J.; Goble, S.; Perrotto, N.; Pinto, S.; Shen, D.-M.; Nargund, R.; Balkovec, J.; DeVita, R. J.; Dreher, S. D. Microscale High-Throughput Experimentation as an Enabling Technology in Drug Discovery: Application in the Discovery of (Piperidinyl)Pyridinyl-1 *H* -Benzimidazole Diacylglycerol Acyltransferase 1 Inhibitors. *J. Med. Chem.* **2017**, *60* (9), 3594–3605. https://doi.org/10.1021/acs.jmedchem.6b01543.

(3) Gesmundo, N. J.; Sauvagnat, B.; Curran, P. J.; Richards, M. P.; Andrews, C. L.; Dandliker, P. J.; Cernak, T. Nanoscale Synthesis and Affinity Ranking. *Nature* **2018**, *557* (7704), 228–232. https://doi.org/10.1038/s41586-018-0056-8.

(4) Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis. *Acc. Chem. Res.* **2017**, *50* (12), 2976–2985. https://doi.org/10.1021/acs.accounts.7b00428.

(5) Gesmundo, N.; Dykstra, K.; Douthwaite, J. L.; Kao, Y.-T.; Zhao, R.; Mahjour, B.; Ferguson, R.; Dreher, S.; Sauvagnat, B.; Saurí, J.; Cernak, T. Miniaturization of Popular Reactions from the Medicinal Chemists' Toolbox for Ultrahigh-Throughput Experimentation. *Nat. Synth.* **2023**, *2* (11), 1082–1091. https://doi.org/10.1038/s44160-023-00351-1.

(6) Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. Nanomole-Scale High-Throughput Chemistry for the Synthesis of Complex Molecules. *Science* **2015**, *347* (6217), 49–53. https://doi.org/10.1126/science.1259203.

(7) Uehling, M. R.; King, R. P.; Krska, S. W.; Cernak, T.; Buchwald, S. L. Pharmaceutical Diversification via Palladium Oxidative Addition Complexes. *Science* **2019**, *363* (6425), 405–408. https://doi.org/10.1126/science.aac6153.

(8) Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.* **2017**, *8* (6). https://doi.org/10.1021/acsmedchemlett.7b00165.

(9) Shen, Y.; Mahjour, B.; Cernak, T. Development of Copper-Catalyzed Deaminative Esterification Using High-Throughput Experimentation. *Commun. Chem.* **2022**, *5* (1), 83. https://doi.org/10.1038/s42004-022-00698-0.

(10) Douthwaite, J. L.; Zhao, R.; Shim, E.; Mahjour, B.; Zimmerman, P. M.; Cernak, T. Formal Cross-Coupling of Amines and Carboxylic Acids to Form Sp3–Sp2 Carbon–Carbon Bonds. *J. Am. Chem. Soc.* **2023**, *145* (20), 10930–10937. https://doi.org/10.1021/jacs.2c11563.

(11) Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T.; Davies, I. W.; DiRocco, D. A.; Sheng, H.; Welch, C. J.; Dreher, S. D. Mapping the Dark Space of Chemical Reactions with Extended Nanomole Synthesis and MALDI-TOF MS. *Science* **2018**, *361* (6402), eaar6236. https://doi.org/10.1126/science.aar6236.

(12) Mahjour, B.; Shen, Y.; Cernak, T. Ultrahigh-Throughput Experimentation for Information-Rich Chemical Synthesis. *Acc. Chem. Res.* **2021**, *54* (10). https://doi.org/10.1021/acs.accounts.1c00119.

(13) Wong, H.; Cernak, T. Reaction Miniaturization in Eco-Friendly Solvents. *Curr. Opin. Green Sustain. Chem.* **2018**, *11*, 91–98. https://doi.org/10.1016/j.cogsc.2018.06.001.

(14) Żurański, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54* (8), 1856–1865. https://doi.org/10.1021/acs.accounts.0c00770.

(15) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOlit, a Small-Size Literature Data Set of Nickel Catalyzed C–O Couplings. *J. Am. Chem. Soc.* **2022**, *144* (32), 14722–14730. https://doi.org/10.1021/jacs.2c05302.

(16) Stevens, J. M.; Li, J.; Simmons, E. M.; Wisniewski, S. R.; DiSomma, S.; Fraunhoffer, K. J.; Geng, P.; Hao, B.; Jackson, E. W. Advancing Base Metal Catalysis through Data Science: Insight and Predictive Models for Ni-Catalyzed Borylation through Supervised Machine Learning. *Organometallics* **2022**, *41* (14), 1847–1864. https://doi.org/10.1021/acs.organomet.2c00089.

(17) Lexa, K. W.; Belyk, K. M.; Henle, J.; Xiang, B.; Sheridan, R. P.; Denmark, S. E.; Ruck, R. T.; Sherer, E. C. Application of Machine Learning and Reaction Optimization for the Iterative Improvement of Enantioselectivity of Cinchona-Derived Phase Transfer Catalysts. *Org. Process Res. Dev.* **2022**, *26* (3), 670–682. https://doi.org/10.1021/acs.oprd.1c00155.

(18) Mahjour, B.; Zhang, R.; Shen, Y.; McGrath, A.; Zhao, R.; Mohamed, O. G.; Lin, Y.; Zhang, Z.; Douthwaite, J. L.; Tripathi, A.; Cernak, T. Rapid Planning and Analysis of High-Throughput Experiment Arrays for Reaction Discovery. *Nat. Commun.* **2023**, *14* (1), 3924. https://doi.org/10.1038/s41467-023-39531-0.

(19) Xbox Controller Clipart. https://www.pikpng.com/pngvi/hJwiiw_xbox-clipart-ps4-controller-xbox-controller-template-png-download/.

(20) Cook, A.; Clément, R.; Newman, S. G. Reaction Screening in Multiwell Plates: High-Throughput Optimization of a Buchwald–Hartwig Amination. *Nat. Protoc.* **2021**, *16* (2), 1152–1169. https://doi.org/10.1038/s41596-020-00452-7.

(21) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**, *583* (7815), 237–241. https://doi.org/10.1038/s41586-020-2442-2.

Appendices

Appendix A: Supplementary Material for Chapter 2

A.1 Code and data availability

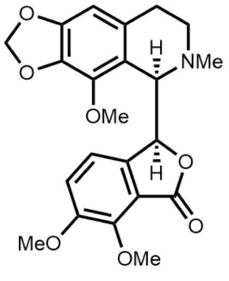
All code for matrix generation, molecular property computation, and figure plotting are available at https://github.com/cernaklab/acid-amine-enumeration-2. A demo of the same script be found at https://github.com/cernaklab/acid-amine-enumeration-2/tree/main/Demo. This folder contains the same hierarchy of script and directory paths as the main code, with only a few modifications to run on a smaller set of matrices and only utilizing one core. Drug structures were accessed from the Drugbank database via https://go.drugbank.com/releases, version 5.1.8.

A.2 Computational methods

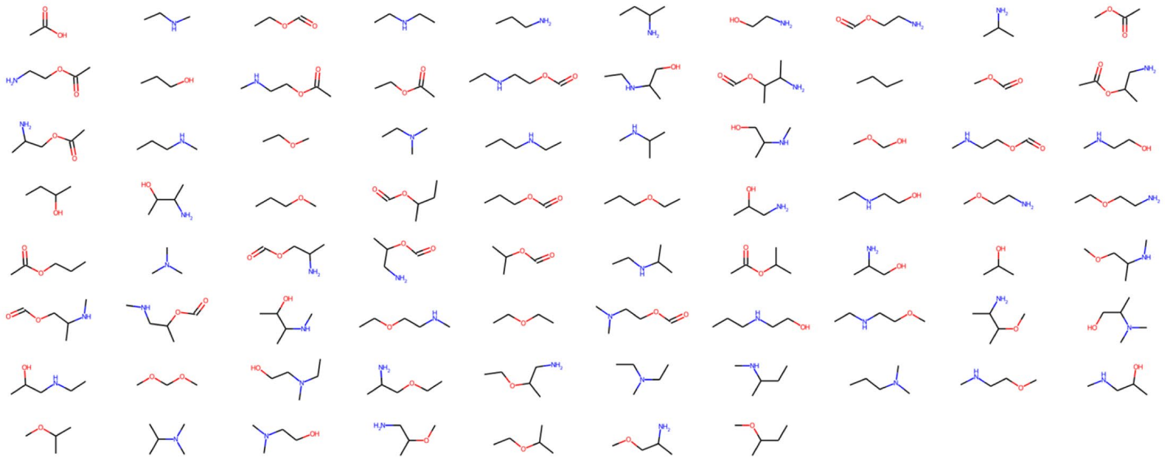
All computation for the manuscript was performed in a Conda environment with the following packages: ipython 7.16.1, jupyterlab 3.1.4, matplotlib 3.3.4, numpy 1.19.2, pandas 1.1.3, RDKit 2019.09.379, seaborn 0.11.1, umap-learn 0.5.1, circos 0.69-9. All packages except Circos are installed via conda-forge or pip. Chord diagrams were plotted using the Circos package, downloaded via http://circos.ca/software/download/circos/.

The Openeye Toolkit 2021.2.0 was used for the docking protocols. Conformations were generated using OMEGA with default torsion libraries after selecting a reasonable tautomer and then performing stereo expansion for undefined stereo centers. The receptors for each protein were docked using FRED with default parameters.

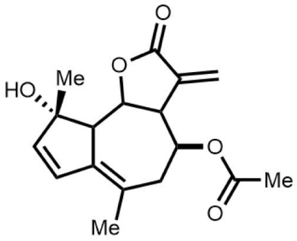
A.3 List of amine–carboxylic acid coupling products present as substructures in drugs



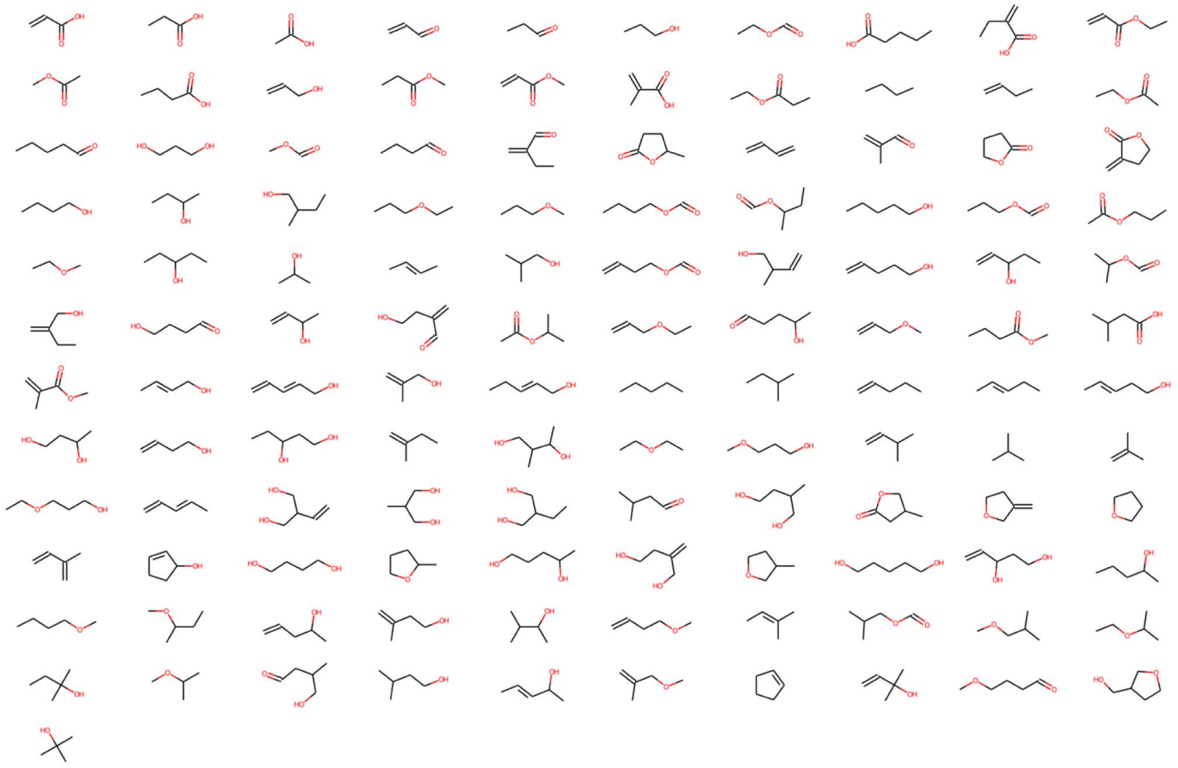
2.25



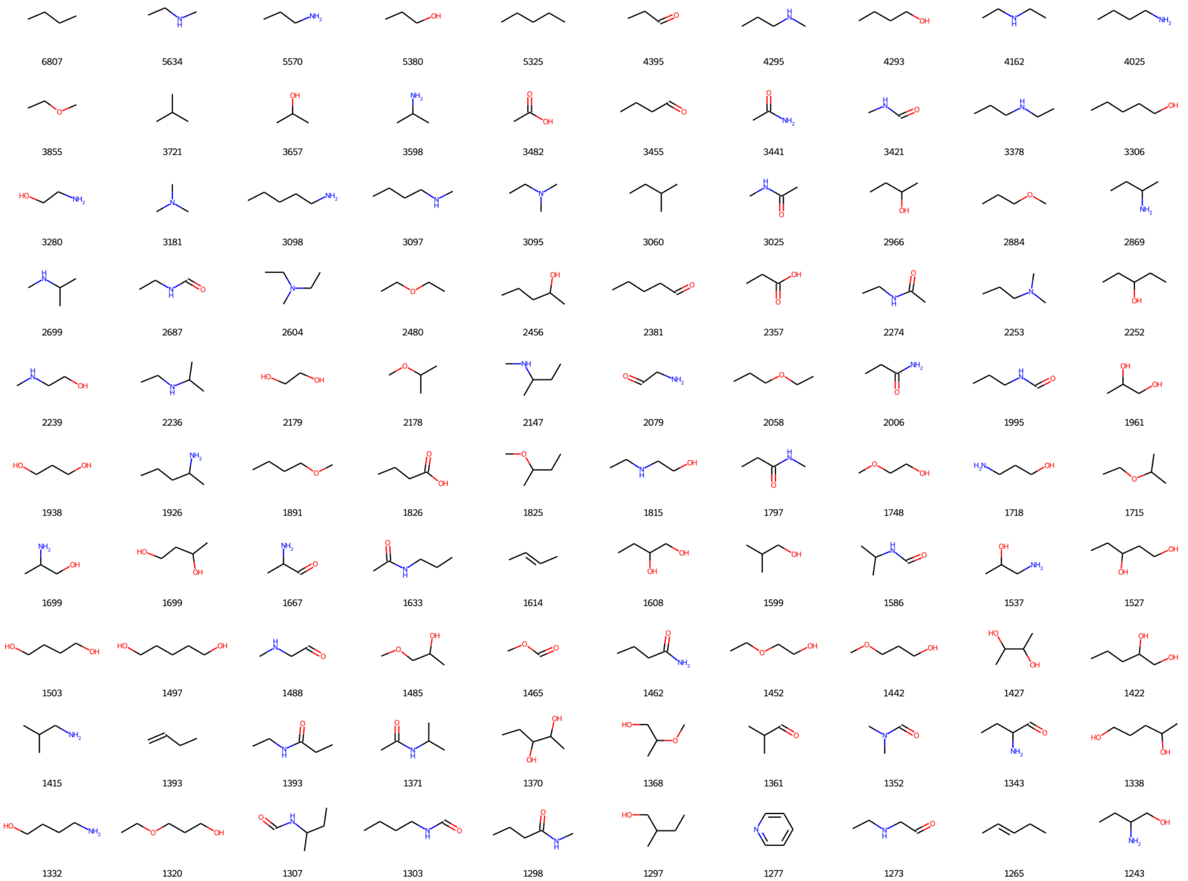
Appendix Figure A.1. All amine–acid reaction products in this work found in noscapine. The structures are arrayed in increasing minimum graph edit distance from a simple 2-carbon amine and 3-carbon acid pair.



2.30



Appendix Figure A.2. All amine–acid reaction products in this work found in athamontanolide. The structures are arrayed in increasing minimum graph edit distance from a simple 2-carbon amine and 3-carbon acid pair.



Appendix Figure A.3. Top 100 most frequently occurring amine–acid enumeration products found as substructures in Drugbank, labeled by number of drugs each product is found in.

Appendix B: Supplementary Material for Chapter 3

B.1 Code availability

Code for graph editing techniques applied towards synthesis of stemoamide (Chapter 3.3) can be found at https://github.com/cernaklab/synthetic-key-steps.

Code for all other sections of Chapter 3 can be found at https://github.com/cernaklab/RZhang_Thesis/.

SYNTHIA™ is available at www.synthiaonline.com.

B.2 Computational environment

All code is written and executed with python version 3.9.7. All dependencies are installed using conda (version 4.10.3). Versions of selected packages are as follows: ipython 7.27.0, jupyterlab 3.1.12, matplotlib 3.4.3, numpy 1.20.3, pandas 1.3.3, RDKit 2021.03.5. A full list of all package versions in the conda environment used is downloadable at https://github.com/cernaklab/synthetic-key-steps/blob/main/software_versions.txt.

B.3 Procedure for Generation of Matrix-Encoded Synthetic Routes

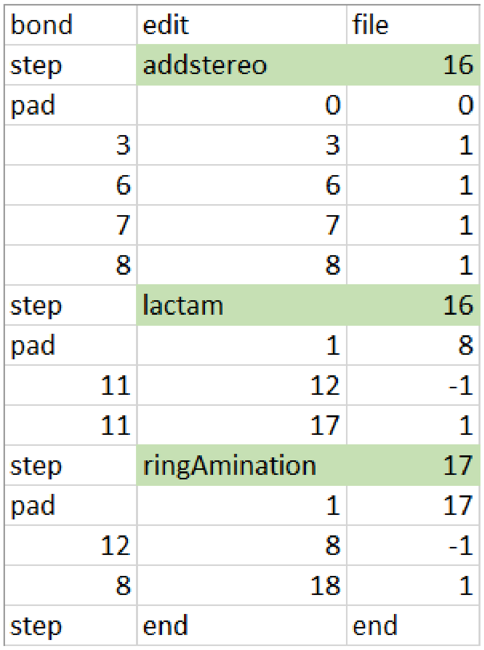
First, adjacency matrices for the target molecule and all synthetic intermediates are generated. Each resultant matrix is an $(N_t + N_c) \times (N_t + N_c)$ square, where N_t is the number of atoms and groups in the target molecule, and N_c is the number of concession atoms and groups. A group refers to a collection of atoms that do not undergo any transformation throughout the synthetic route, and hence can be encoded as a single atom without losing information relevant to the results presented in this manuscript. Examples include alkyl side chains and protecting groups.

Several methods are possible for the generation of these matrices. The process employed by this manuscript is as follows:

1. Encode the product as a mol object in RDKit, replacing large unchanged substituents with single atoms as desired.

2. Iterate over the product’s atoms using GetAtoms(), and extract their atomic number using GetAtomicNum(). Should the mol object be encoded directly from a SMILES string, the order in which atoms appear in GetAtoms()will be the same as that in the SMILES string.
3. Use GetAdjacencyMatrix(), setting useBO=True to generate the product adjacency matrix from the mol object. The ordering of product atoms along the rows and columns will be the same as that extracted in step 2.
4. Working backwards from the product, note down the following for every reverse step:
    a. The number and identities of additional concession atoms to be appended.
    b. Atom indices at which there is a change in bond order.
    c. Changes in symmetry about atoms that are asymmetric in the product.

These changes are transferred to a bond edit spreadsheet. An example of a short spreadsheet is depicted in Figure B1.



bond	edit	file
step	addstereo	16
pad	0	0
3	3	1
6	6	1
7	7	1
8	8	1
step	lactam	16
pad	1	8
11	12	-1
11	17	1
step	ringAmination	17
pad	1	17
12	8	-1
8	18	1
step	end	end

Appendix Figure B.1. A sample bond edit spreadsheet, with which the matrix-encoded synthetic route is generated.

Each retrosynthetic step is split into three sections:

1. The “step” row, which separates individual steps. The cells highlighted in green are not read by the matrix generation algorithm, and are intended as flexible annotation spaces for

the user. In the files for this manuscript, they are most commonly used to keep track of step names and maximum atom count.

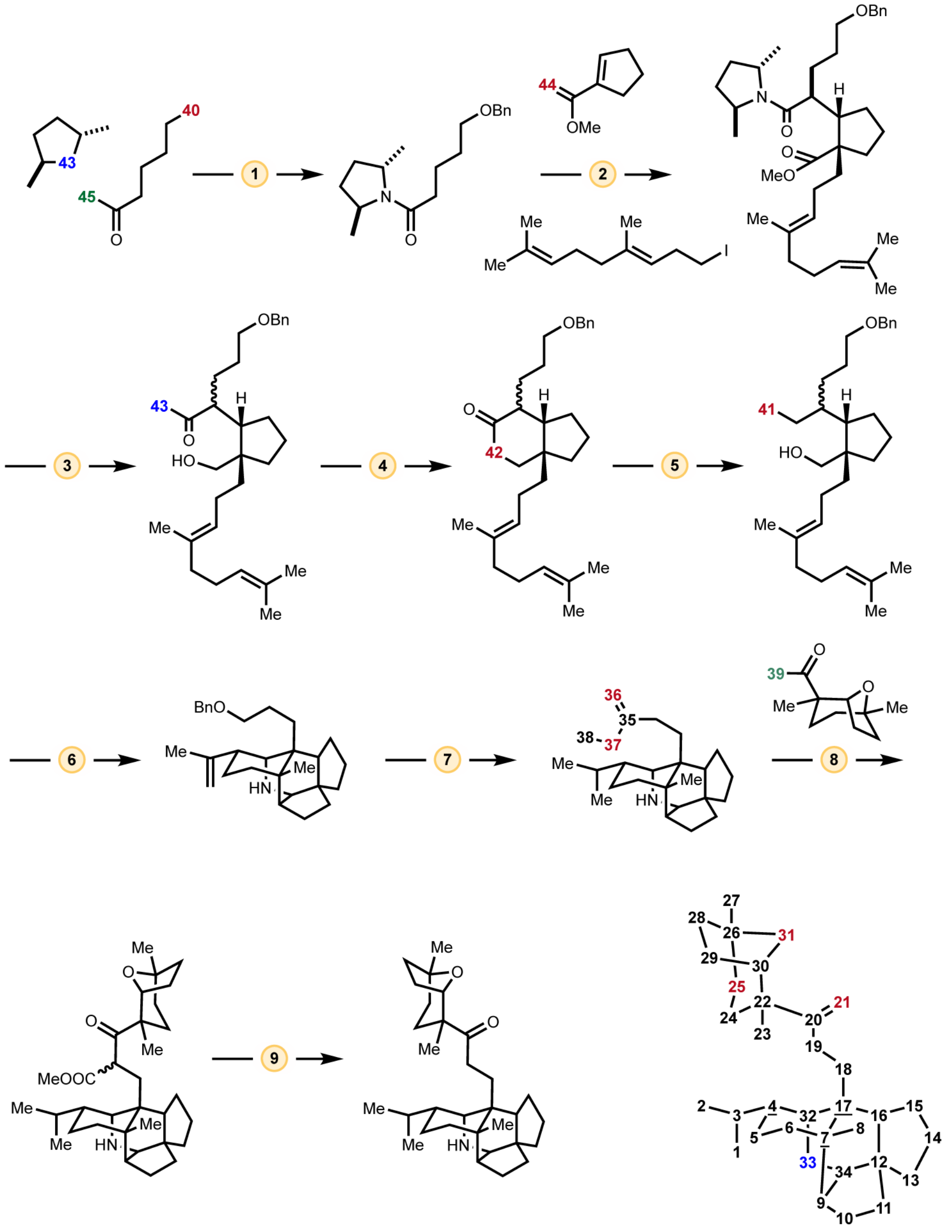
2. The “pad” row, whose first entry is the number of concessions atoms to add, and the second is their atomic numbers, in order, separated by spaces.
3. The bond edit block, where, for each row, the first two entries are indices of the atoms joined by a changing bond, and the third is the change in bond order.

After the final step is encoded (i.e. the first step in the forward synthesis), a last row is added to mark the end of the bond edit file.

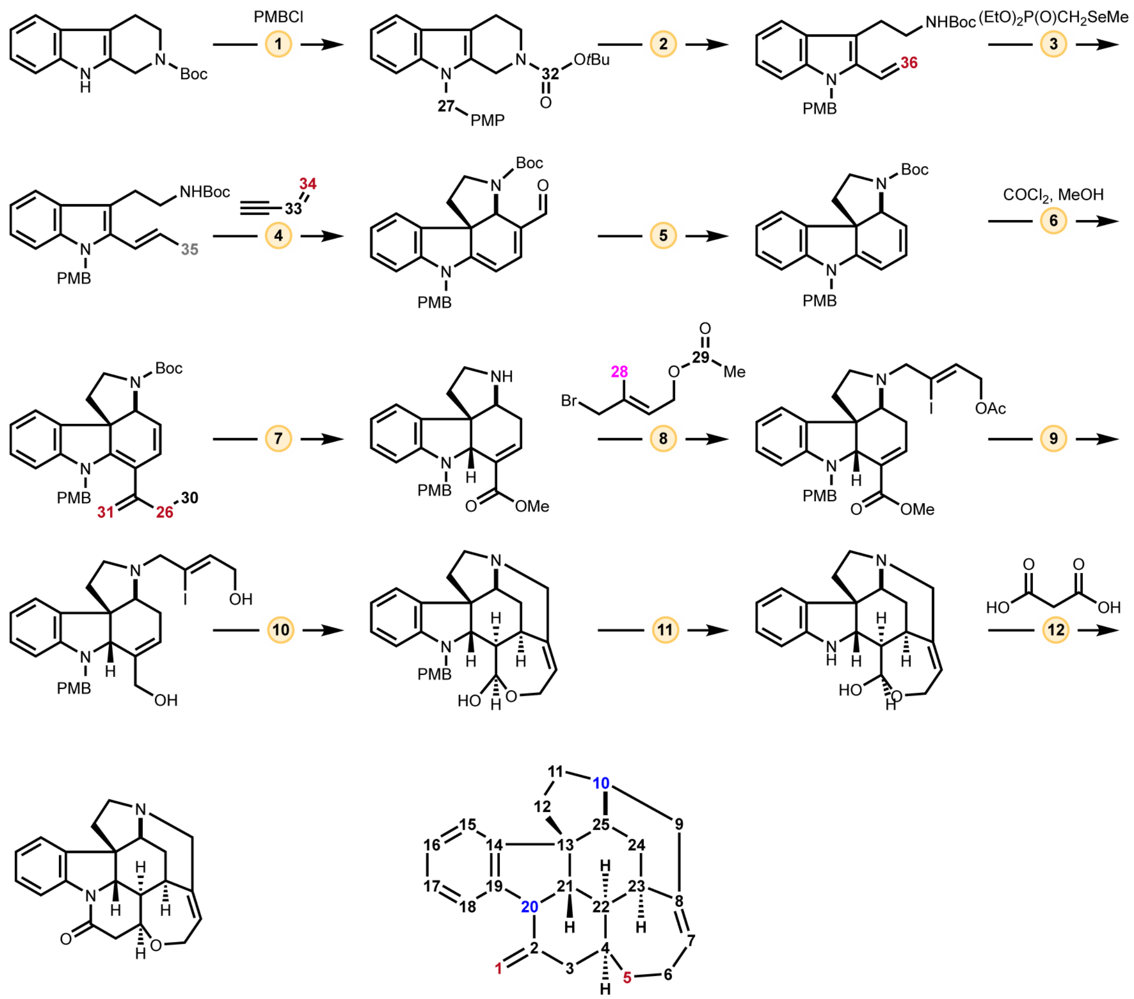
Most edit files for this manuscript use the first “step” to add in stereocenters as diagonal entries. R and S stereocenters are not distinguished in this work, and only encoded as 1 for present, and 0 for absent. Stereocenters that appearing during the synthesis, but are absent in the final product, are not encoded.

After the bond edit file is complete, it is used to compute and save the adjacency matrices for the remaining intermediates. Should any changes be difficult to encode via a bond edit spreadsheet, they can be added to the matrices after they are generated.

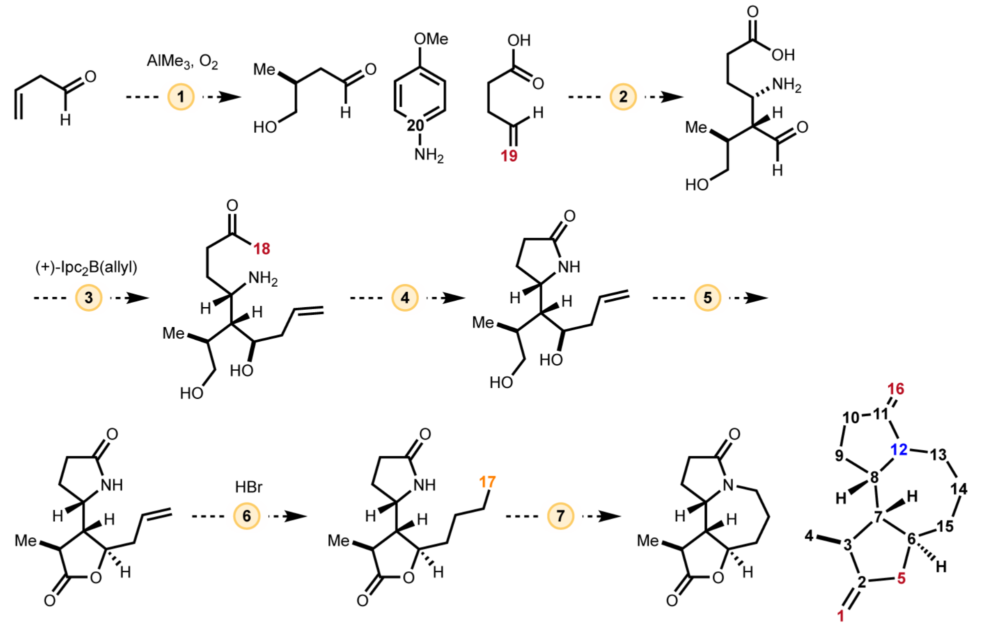
B.4 Schematics and atom indices of all analyzed synthetic routes.



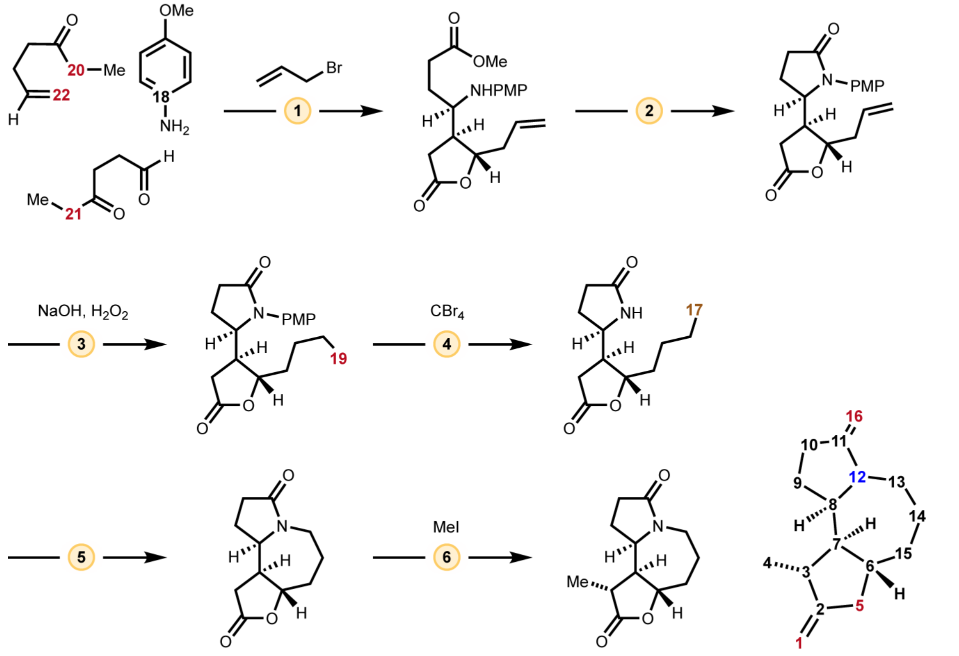
Appendix Figure B.2. Total synthesis of (–)-secodaphniphylline by Heathcock and Stafford, including atom mapping.



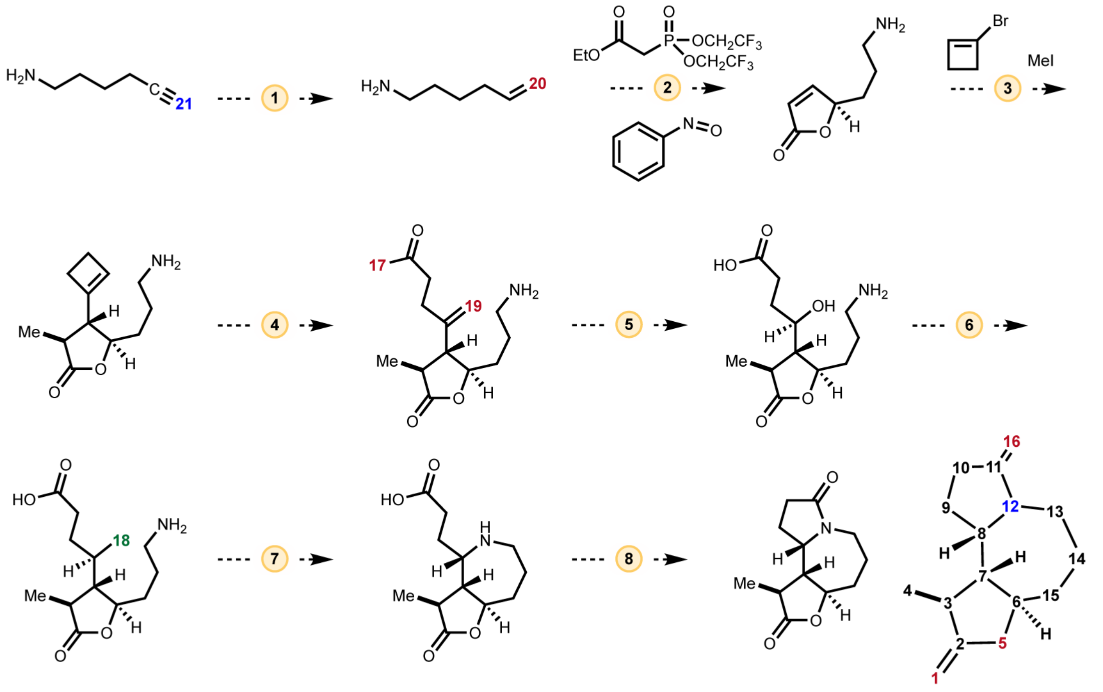
Appendix Figure B.3. Total synthesis of (–)-strychnine by MacMillan and coworkers, including atom mapping.



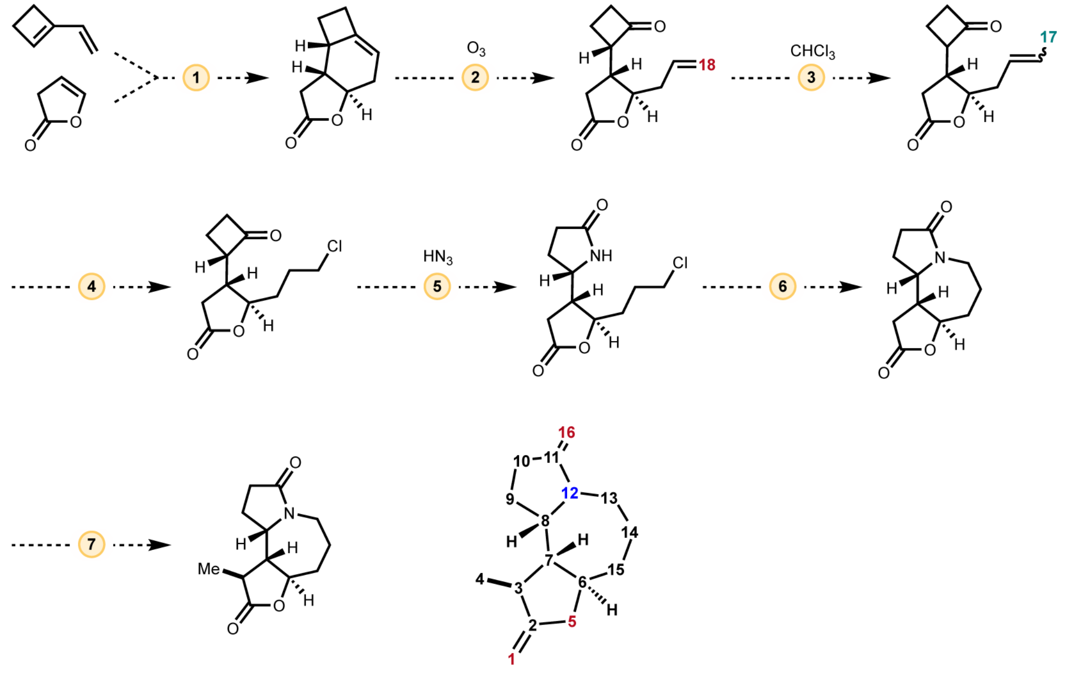
Appendix Figure B.4. Total synthesis route towards (–)-stemoamide as proposed by the CASP program SYNTHIA™, including atom mapping.



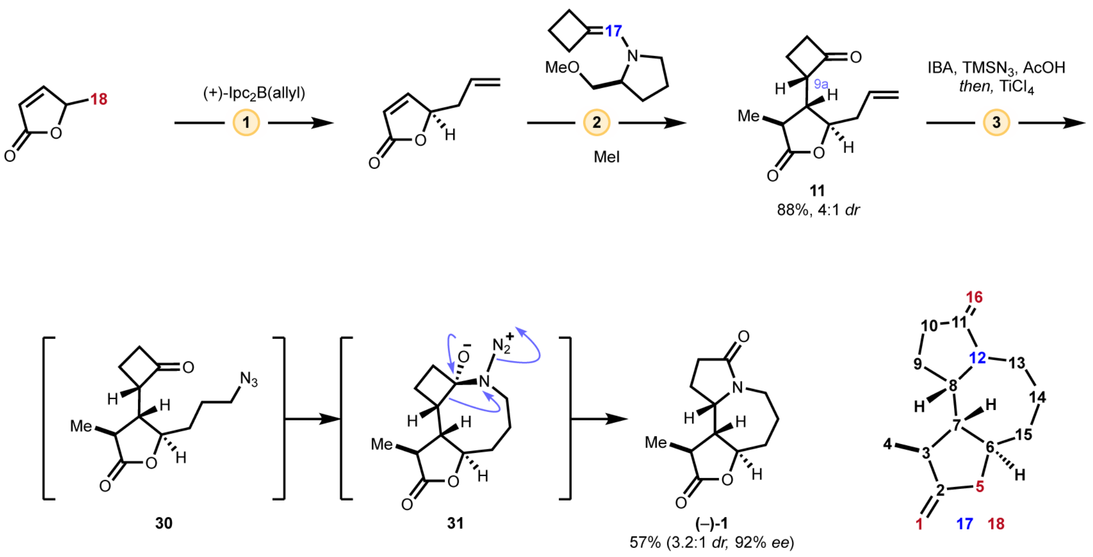
Appendix Figure B.5. Six-step total synthesis of (+)-stemoamide performed by our group, including atom mapping.



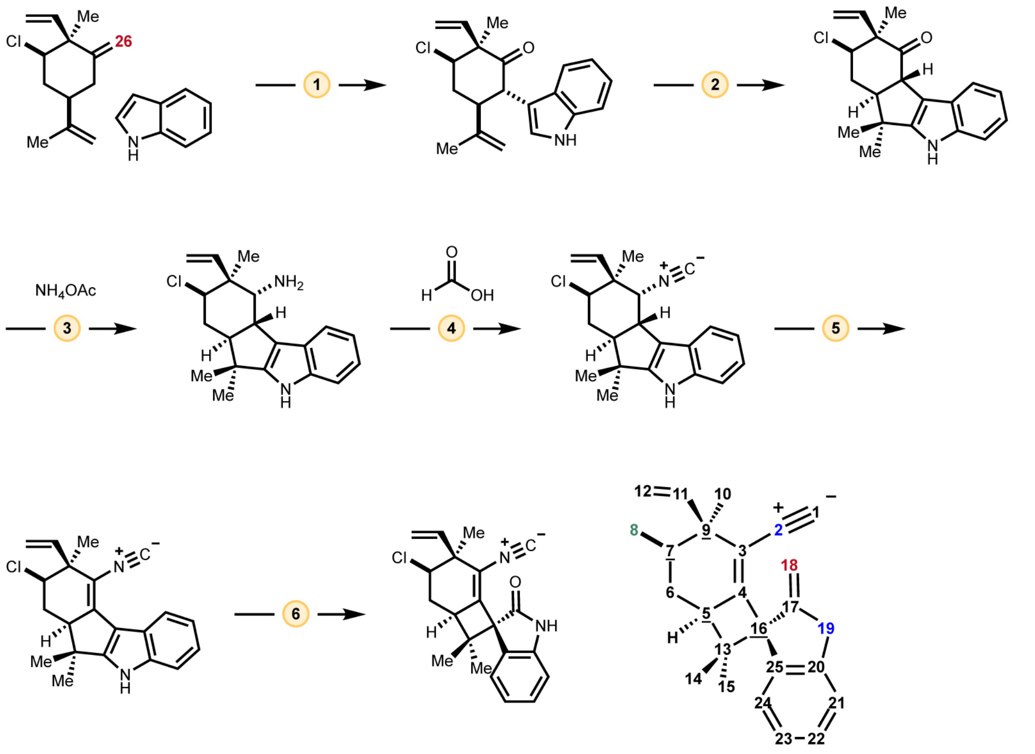
Appendix Figure B.6. Total synthesis route towards (–)-stemoamide as proposed by SYNTHIA™ with the Mannich reaction excluded, including atom mapping.



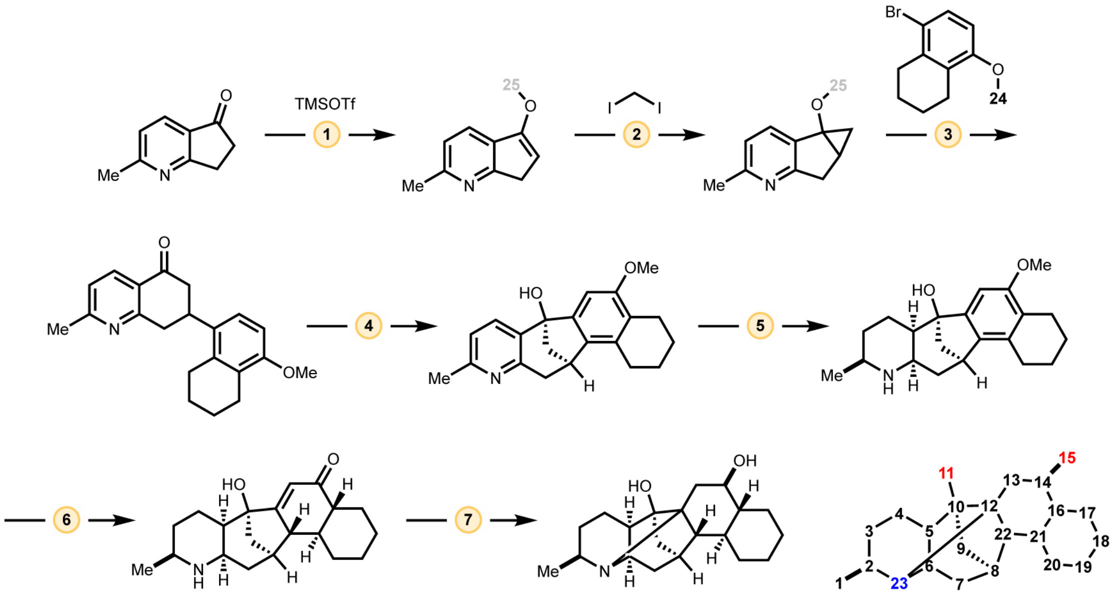
Appendix Figure B.7. Total synthesis route towards (–)-stemoamide as proposed by SYNTHIA™ when the penultimate intermediate is used as the target, including atom mapping.



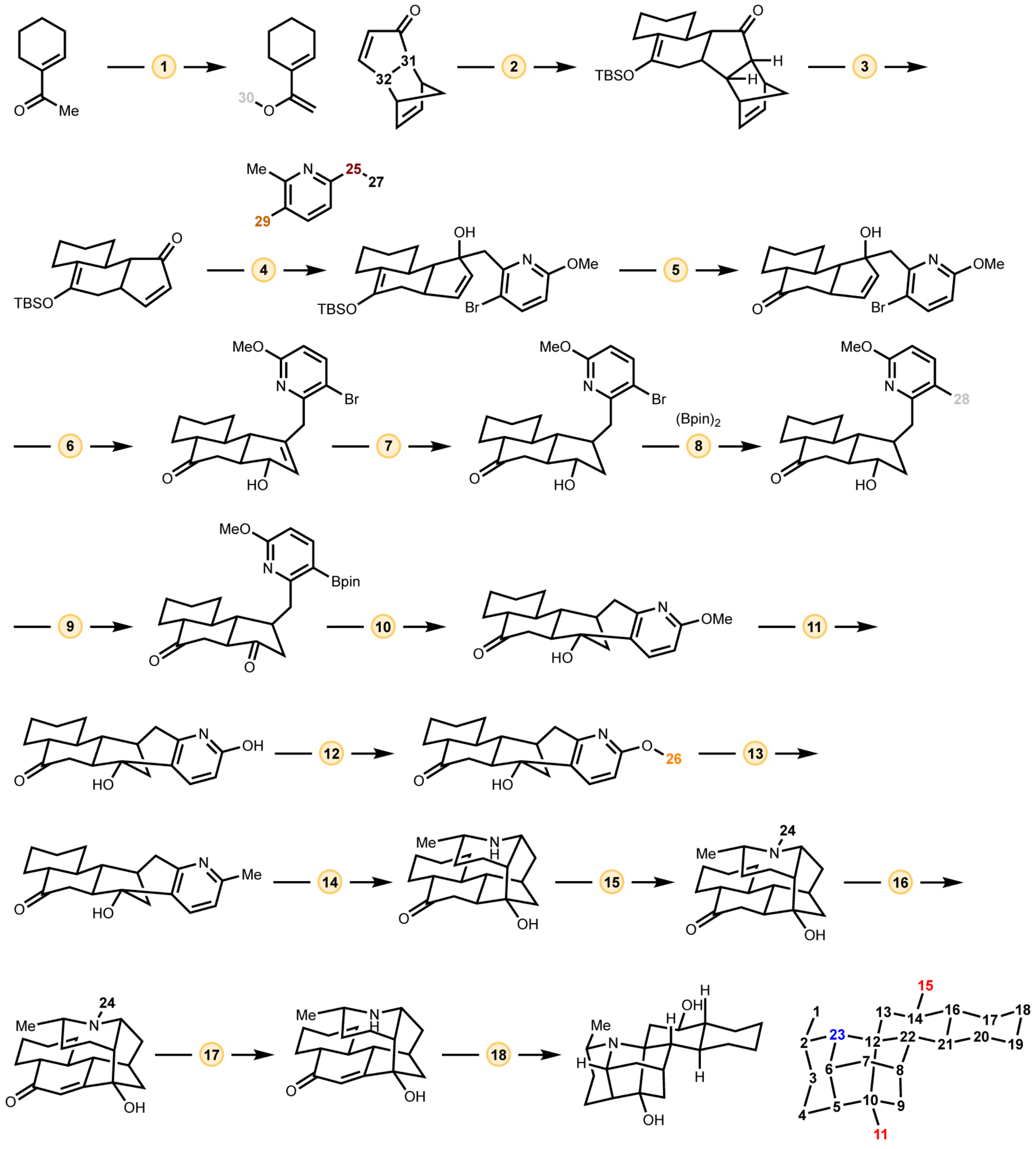
Appendix Figure B.8. Three-step total synthesis of (+)-stemoamide performed by our group, including atom mapping.



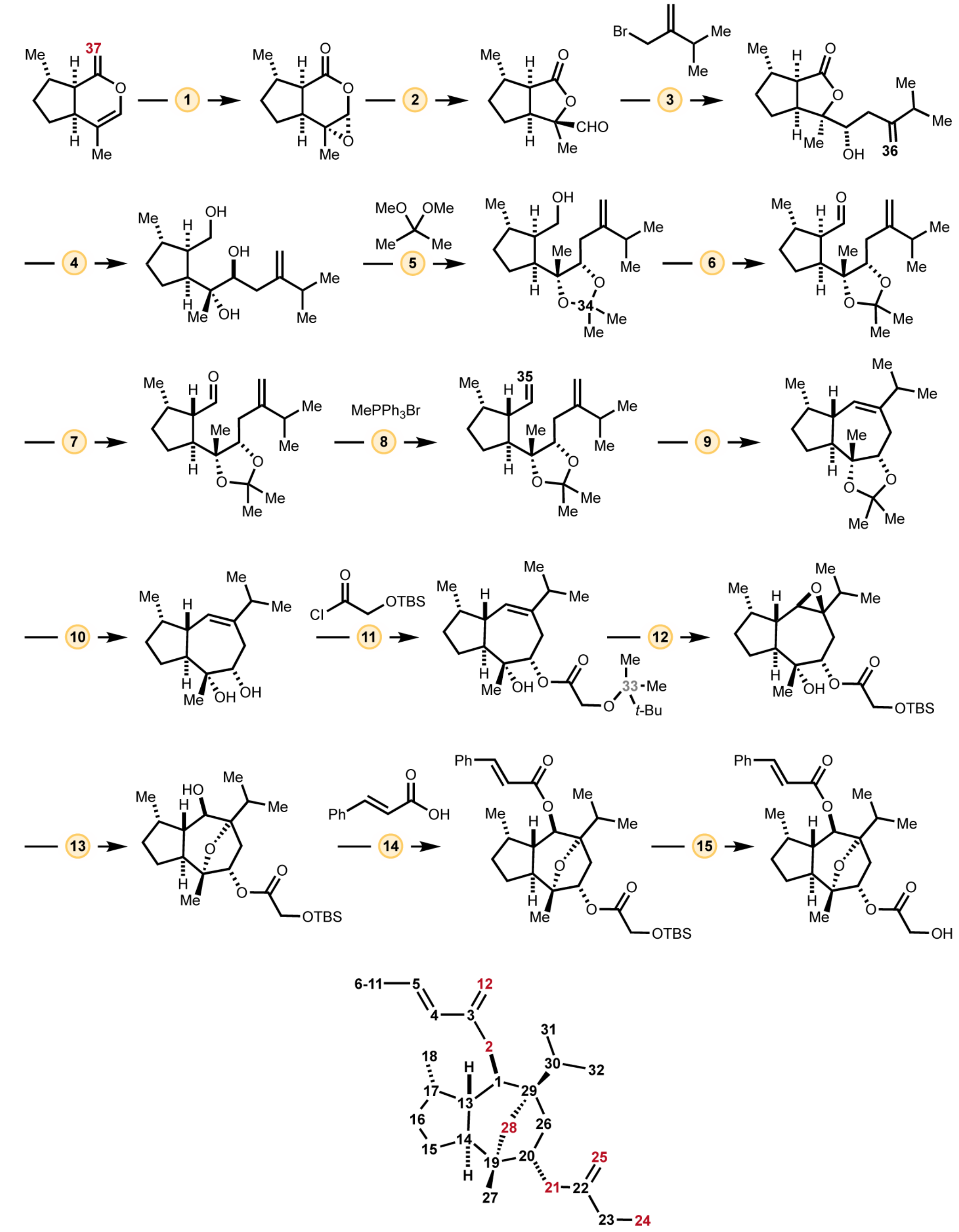
Appendix Figure B.9. Total synthesis of (+)-welwitindolinone A by Baran and coworkers, including atom mapping.



Appendix Figure B.10. Total synthesis of himgaline by Shenvi and coworkers, including atom mapping.



Appendix Figure B.11. Total synthesis of himgaline by Sarpong and coworkers, including atom mapping.



Appendix Figure B.12. Total synthesis of englerin A by Christmann and coworkers, including atom mapping.

Appendix C: Supplementary Material for Chapter 4

C.1 Code availability

Code for small-scale library synthesis (Chapter 4.4.2) can be found at
https://github.com/cernaklab/McGrathEsterification.
Code for all other sections of Chapter 4 can be found at
https://github.com/cernaklab/RZhang_Thesis.
Software for controller mapping to keyboard inputs was downloaded from
https://github.com/AntiMicro/antimicro.

C.2 General procedure for conducting HTE screens

All reagents to be screened were entered into phactor™ either manually, or selected from the online inventory. The molar mass, reagent type, molarity in the reaction vessel, amount of overage, and order of addition were entered for each reagent, and a recipe is automatically generated.

Following the recipe, all reagents were measured into oven-dried glass vials (ChemGlass #CG-4912-02) by mass or volume. A magnetic stir bar (Fisher Scientific 14-513-93) is placed in all vials where the reagent is not expected to be completely soluble at the stock solution concentration, unless the reagent is to be dosed as a solid.

A 96-well aluminum microvial plate (Analytical Sales & Services cat. no. 25243) was equipped with oven-dried shell vials (Analytical Sales & Services cat. no. 884001) and one parylene-coated stir dowel (Analytical Sales & Services cat. no. 13258) was placed in each vial.
All labware and reagents were brought into a glovebox (MBraun LABmaster Pro) and onto the deck of an Opentrons OT-2, fitted with a tumble stirrer (V&P Scientific Inc. 710D3). Reagents that require stirring are placed on a 24-well stirring block (Analytical Sales & Services #24125) which fits on the tumble stirrer deck. Stock solutions are prepared by adding the prescribed amount of solvent into each reagent vial, either manually with a single-channel micropipetter, or a single-channel Opentrons autopipetter dosing from a deep well reservoir (Analytical Sales & Services #

962144) using pipette tips supplied by the manufacturer (Opentrons PT0300-9B-NS). The tumble stirrer was activated to agitate stock solutions while the script to dose stock solutions into the microvial plate is being run. Pauses can be pre-programmed into the Opentrons script to allow for pre-mixing, or addition of solid reagent. After all stock solutions have been dosed, the microvial plate was sealed with two layers of rubber mat (Analytical Sales & Services # 96965) and one layer of PFA film (Analytical Sales & Services # 96979), removed from the glovebox, and heated using a heating block (V&P Scientific Inc. 741GA) for the desired temperature and duration, stirring at 500 RPM (V&P Scientific Inc. 710E5X tumble stirrer).

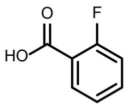
After the reaction time has elapsed, the microvial plate was returned to the robot deck along with a polypropylene 96-well deep well plate (Analytical Sales & Services # 17P687Z) and a fresh deep well reservoir containing caffeine solution in Optima grade acetonitrile in one well, and only acetonitrile in the other. An 8-channel pipette first transfers a calculated amount of the caffeine solution into the microvial plate, and then an aliquot of these mixtures is transferred into the deep well polypropylene plate, with premixing of 3 repetitions of around half the total liquid volume. Lastly, pure acetonitrile is added to every well of the deep well plate such that each well contains a uniform amount of liquid not less than 600 μL. The deep well plate was centrifuged for 40 minutes at 1000 RPM (Genevac HT-4X) and analyzed using UPLC-MS.

The UPLC-MS was a Waters I-class ACQUITY (Waters Corporation, Milford, MA, USA) equipped with in-line photodiode array detector (PDA) and QDa mass detector (ESI positive ionization mode). 0.1 μL sample injections were taken from acetonitrile solutions of reaction mixtures or products (~1 mg/mL). A partial loop injection mode was used with the needle placement at 1.0 mm from bottom of the wells and a 0.2 μL air gap at pre-aspiration and post-aspiration. Column used: Waters Cortecs UPLC C18+ column, 2.1mm · 50 mm with (Waters #186007114) with Waters Cortecs UPLC C18+ VanGuard Pre-column 2.1mm · 5 mm (Waters #186007125), Mobile Phase A: 0.1 % formic acid in Optima LC/MS-grade water, Mobile Phase B: 0.1% formic acid in Optima LC/MS-grade MeCN. Flow rate: 0.8 mL/min. Column temperature: 45 °C. The PDA sampling rate was 20 points/sec. The QDa detector monitored m/z 150-750 with a scan time of 0.06 seconds and a cone voltage of 30 V. The PDA detector range was between 210 nm – 400 nm with a resolution of 1.2 nm. 1-minute and 2-minute methods were used. The method gradients are as follows: 0 min: 0.8 mL/ min, 95% 0.1% formic acid in water/ 5% 0.1% formic acid in acetonitrile; 1.5 min: 0.8 mL/ min, 0.1% 0.1% formic acid in water/ 99.9% 0.1% formic acid in acetonitrile; 1.91 min: 0.8 mL/min, 95% 0.1% formic acid in water/ 5% 0.1% formic acid in acetonitrile.

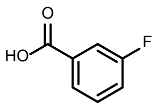
In experiments covered in Chapter 4.3, the caffeine solution was at a concentration of 16 mg/mL, 500 µL of which was transferred into each well. An aliquot of 70 µL was transferred into the deep well plate, and 730 µL of pure acetonitrile was added.

In experiments covered in Chapter 4.4, the caffeine solution was at a concentration of 0.1 M, 100 µL of which was transferred into each well. An aliquot of 40 µL was transferred into the deep well plate, and 560 µL of pure acetonitrile was added.

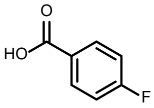
C.3 Structures of Carboxylic Acids Used in Small Scale Library Synthesis (Chapter 4.4.2)



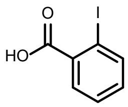
A1



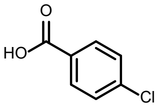
A2



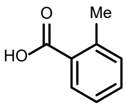
A3



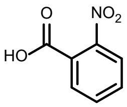
A4



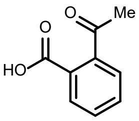
A5



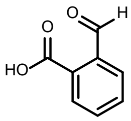
A6



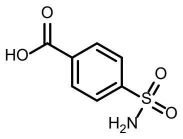
A7



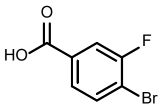
A8



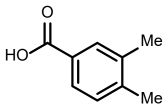
A9



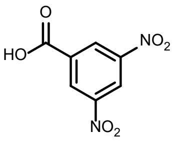
A10



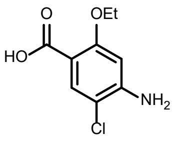
A11



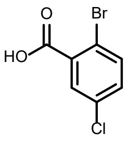
A12



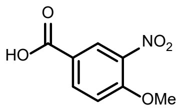
B1



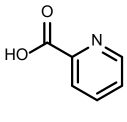
B2



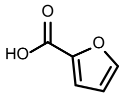
B3



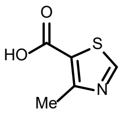
B4



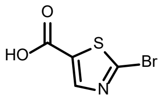
B5



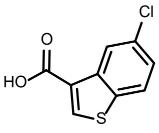
B6



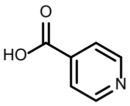
B7



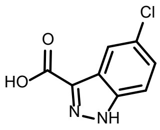
B8



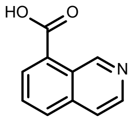
B9



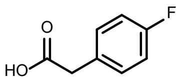
B10



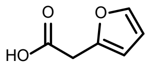
B11



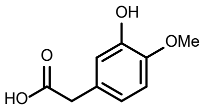
B12



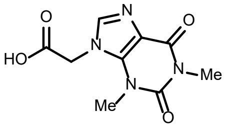
C1



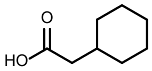
C2



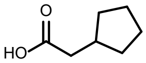
C3



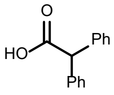
C4



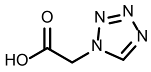
C5



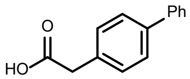
C6



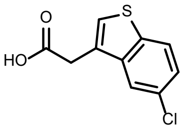
C7



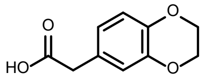
C8



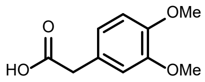
C9



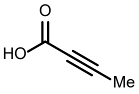
C10



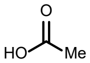
C11



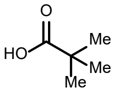
C12



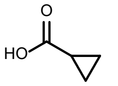
D1



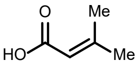
D2



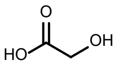
D3



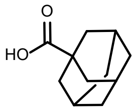
D4



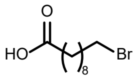
D5



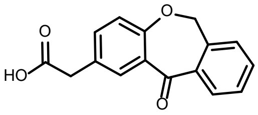
D6



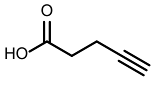
D7



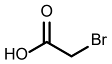
D8



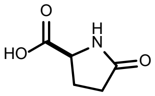
D9



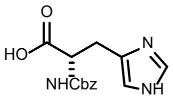
D10



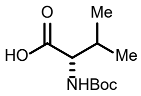
D11



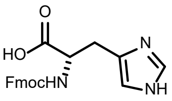
D12



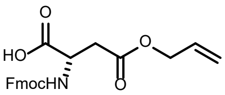
E1



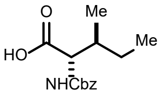
E2



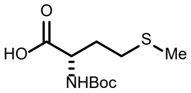
E3



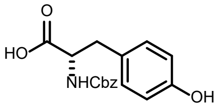
E4



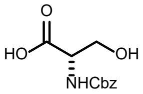
E5



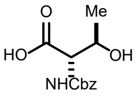
E6



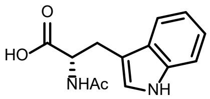
E7



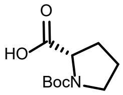
E8



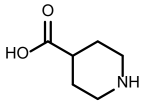
E9



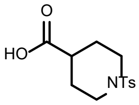
E10



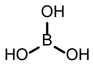
E11



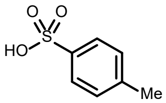
E12



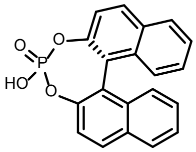
F1



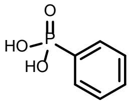
F2



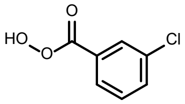
F3



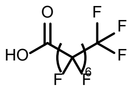
F4



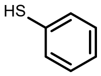
F5



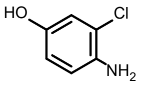
F6



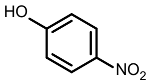
F7



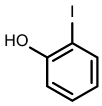
F8



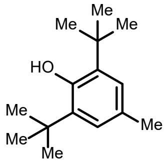
F9



F10



F11



F12