**One-Step Retrosynthesis of Drug Molecules Leveraging C–H Coupling Reactions with Commercially Available Building Blocks**

Babak Mahjour, Kaitlyn M. Flynn, Shannon Stahl\*, Tim Cernak\*

1. Department of Medicinal Chemistry, University of Michigan, Ann Arbor, MI, USA
2. Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA

**Abstract.** Synthetic access to drugs and other complex molecules relies on the existence of diverse retrosynthetic disconnections, made possible through the availability of a modern toolbox of chemical reaction methods. As new reaction methods are developed, multistep synthesis routes can be increasingly streamlined since novel bond disconnections become available through the development of novel reactivities. Here, we map all possible one-step retrosyntheses of drugs from the DrugBank database to commercially available building blocks from a vendor catalog, considering all routes whether the proposed disconnection invokes a reaction method that is already known or has not yet been reported. The popularity of C–H bonds in commercially available building blocks highlights that developing new site selective C–H bond functionalization reactions could provide access to many drugs in a single step from commercial materials. These one-step retrosyntheses could impact both the process scale-up of single drug targets, and the discovery exploration of diverse chemical space around medicinal chemistry lead compounds.

**Introduction**

Modern computer-aided synthesis enables efficient identification of retrosynthetic routes to small molecule targets.1 While this approach can facilitate the design of complex molecule synthesis, algorithms generally rely on established reaction methods.2 Enumerative combinatorics is a technique that can explore all possible bond formations for a given pair of building blocks (Figure 1A). For example, this approach has been used to demonstrate the array of coupling products accessible from amines and carboxylic acids beyond amides.3 This enumerative strategy does not require that each coupling reaction to map onto a known synthetic method, and also includes underdeveloped or unknown transformations that could be envision from the pair of building blocks. As this approach is unbiased by existing reaction methods, it can point to opportunities for method development. Here, we show how this approach may be adapted to retrosynthesis of drug molecules, with the goal of identifying opportunities for reaction development that could provide efficient access to drug molecules and analogs thereof. The retrosynthetic method outlined herein is constrained to one-step disconnection of targets from DrugBank compounds into commercially available building block synthons. The results highlight the abundance of activated C(*sp*3)–H building blocks and illustrates how new reaction methods could provide efficient direct access to existing drug molecules and their analogs.

Our study takes advantage of two available databases, including drug compounds from the DrugBank database,4 and commercially accessible compounds through the MilliporeSigma catalog. An example retrosynthesis of a DrugBank compound to commercially available building block synthons is shown in Figure 1B. Through a single bond formation, it is feasible drug **1** shown can be directly accessed from three different sets of commercially available compounds – **2** + **5**, **3** + **6**, or **4** + **7** – where the bond formed either includes or excludes the existing building block functional handle. A facile way to visually represent a vast array of viable retrosyntheses of DrugBank compounds is through chord diagrams (Figure 1B, center) where one synthon (Building Block A) is arrayed along the bottom-left arc, the other synthon (Building Block B) is arrayed along the bottom-right arc, and target molecules are arrayed along the top arc. A chord between a synthon and the target molecule indicates that the synthon can be used to form the target in one step when merged with a compound found in the other synthon arc.

Retrosynthetic analysis of 9,082 DrugBank compounds shows that *sp*3–*sp*3 bonds are the most prevalent among these structures, with *sp*2–*sp*2 bonds being least common (*vide infra*). With this in mind, functionalization of *sp*3-hybridized atoms remains a high priority in accessing drugs.

A screenshot of a computer

Description automatically generated

**Figure 1 | Introduction to Enumerative Combinatorics for Retrosynthesis. a)** Enumerative combinatorics for single bond formation between two coupling partners. **b)** Adaptation of enumerative combinatorics for retrosynthetic analysis.

Coupling partners such as halides, boronic acids, alcohols, acids, and amines are all viable using modern synthesis.5 In recent years, activated C–H building blocks have been increasingly utilized for site-selective bond formations. When looking at the relative commercial availability of various building blocks (Figure 2A), it is apparent that alcohols, carboxylic acids and amines are the only common coupling partners that compete with activated C(*sp*3)–H building blocks, such as benzylic C–H motifs, C–H bonds alpha to heteroatoms (i.e. N, O, or S), and enolate precursors with C–H bonds a to a carbonyl group, in terms of availability. Focusing on the (hetero)benzylic C(*sp*3)–H motif, the overlap between commercially available benzylic C–H molecules and their corresponding benzyl amines, halides, and alcohols was compared. In Figure 2B, it is shown that of the ~0.2 million available benzylic C–H compounds, only a fraction of those have corresponding “traditional” cross coupling building blocks, indicating that there is a wealth of unique chemical space that can be accessed through benzylic C–H functionalization.

There has been a substantial expansion of chemistry utilizing diverse classes of activated C–H bonds. In Figure 2C, a few recent examples of benzylic C–H functionalization to form C–N and C–O bonds are shown.6,7 These examples, along with other benzylic transformations developed have been shown to access vast libraries of drug-like compounds. Similar efforts have enabled site-selective functionalization of C–H bonds a to heteroatoms.

A diagram of a complex structure

Description automatically generated with medium confidence

**Figure 2 | C–H motifs represent valuable building blocks for cross coupling. a)** Commercial availability of different coupling partners. **b)** Comparison of commercially available benzylic building blocks. **c)** Examples of utilizing benzylic C–H functionalization to access drug-like compounds.

In this analysis, an algorithm was used to decompose 9,082 drug structures found in DrugBank, which have a molecular weight less than 500 g/mol after desalting, into synthons. In this process, every carbon–carbon, carbon–nitrogen, and carbon–oxygen single bond was deleted and replaced with generic functional group placeholders on both ends of the original bond (cf. A and B in Figure 1A). Subsequently, building block functional handles (acid, alcohol, boronate, amine, iodine, bromine, chlorine, and hydrogen) were enumerated at the disconnection point of both synthons. The full combination of both synthon enumerations was taken, and each pair was cross-referenced against 114,300 commercial compounds found in MilliporeSigma’s catalog with molecular weight less than 200 g/mol. If both synthons were found to be purchasable, the retrosynthetic reaction was saved. Once all possible single step retrosynthetic reactions between drugs and purchasable compounds were recorded, post-hoc refinement was performed to group the reactions into synthetically aware bins. For instance, each transformation containing a non-hydrogen synthon (e.g., deaminative, decarboxylative, etc.) was grouped as aryl or non-aryl (via checking the a-carbon to be equal to a ‘c’ or ‘C’ in SMARTS notation). Synthons containing C–H bonds were analyzed to see if they were benzylic or if the associated carbon was alpha to a carbonyl group or heteroatom, and the remaining, including aryl and unactivated aliphatic hydrogens, were labelled miscellaneous. Finally, to produce the chord diagrams, all transformations containing the functional group to be analyzed were collected, and the synthons containing the functional group in focus were grouped as ‘synthon A’ and the remaining half was grouped as ‘synthon B’.

A screenshot of a cell phone

Description automatically generated

**Figure 3 | Benzylic C–N and C–O Retrosynthetic Analysis. a)** Frequency of occurrence of single bonds in DrugBank molecules. **b)** All hypothetical single bond C–H cross couplings between benzylic carbon atoms and nitrogen atoms to form DrugBank compounds from commercial substrates. **c)** All hypothetical single bond C–H cross couplings between benzylic carbon atoms and oxygen atoms to form DrugBank compounds from commercial substrates.

Upon analyzing the retrosynthetic pathways feasible for benzylic C–N or C–O containing DrugBank compounds to benzylic C–H building blocks, it was revealed that 53 compounds can be directly accessed through benzylic C–N bond formation, and 35 are accessible via benzylic C–O bond formation (Figure 3). Among these 88 disconnections, most of the nitrogen or oxygen motifs would be sourced from the parent amine or alcohol, leading to excellent atom economy, with the remaining disconnections utilizing deaminative and deoxygenative chemistry to accomplish the coupling reaction.

Extending this analysis of benzylic C–H motifs in one-step retrosynthesis, we also analyzed the utility of benzylic alkylation and arylation reactions (Figure 4, center). In total, it was revealed that there are an additional 264 disconnections when extending the study to C–C bond formation, with the new carbon sources coming from a variety of coupling partner classes. Based upon this analysis, it is apparent that there is a lot of value that can be gained from development of site-selective C–H benzylic alkylation or arylation, regardless of the class of coupling partner being used.

The utility for C–C bond formation from other activated C–H building blocks is also addressed in Figure 4. Enolate precursors (α-carbonyl) can be used broadly to access 428 disconnections of DrugBank compounds, many of which are feasible due to the longstanding study of enolate chemistry.

In this analysis, C–H bonds alpha to a heteroatom (N, O, S, etc.) are revealed to be the most promising of these activated C–H bond classes for accessing currently reported DrugBank compounds in one step, with 800 total disconnections accessible from commercially available compounds.

**A chart of different colors

Description automatically generated with medium confidence**

**Figure 4**. **|** **Utilizing Activated C–H Motifs for C–C Bond Cross Coupling.** Analyzing the DrugBank compounds accessible from commercially available building blocks through a one-step alkylation or arylation of an activated C–H bond.

The retrosynthetic analysis previously discussed has shown how useful activated C–H bonds can be in synthesizing drugs. One such showcase of using this analysis to directly access the drug bromhexine (**XX**) is shown in Figure 5A. In this reaction, an activated C–H building block (**XX**) is engaged in an arylation reaction with 2,4,6-tribromoaniline (**XX**) to access the drug in a single step. With increasing methodology, such as this example, being developed, these motifs will also prove to be extremely useful in a medicinal chemistry context for directly accessing libraries of compounds in a high-throughput fashion to facilitate drug discovery. To showcase one example of an accessible virtual library, we enumerated C–H substrate **XX** with a set of 1,000 commercially available aryl bromides (Figure 5B) using the a-amino C–H arylation depicted in Figure 5A. In this virtual library of bromhexine congeners, it is revealed that a wide range of physicochemical properties can be accessed. Depending on the objective, specific property distributions can be targeted to optimize lipophilicity, molecular weight, quantitative estimate of druglikedness (QED) or other physicochemical properties (Figure 5C). When mapped to a *t*-distributed stochastic neighbor embedding (Figure 5D), it is clear that a broad diversity of chemical space can be accessed.

We conclude from this analysis that activated C–H bonds, including those at benzylic sites and a to carbonyls or heteroatoms, make extremely valuable cross-coupling partners. These building blocks can not only be used to reimagine retrosynthesis of reported DrugBank compounds, but also to streamline drug discovery through large library synthesis, due to the high commercial availability of these compounds. This method of retrosynthesis contrasts contemporary generative algorithms which typically require additional information to ensure synthetic feasibility. Emphasis should be placed on development of site-selective cross coupling reactions between these activated C–H motifs and all classes of common coupling partners, to maximize the potential of one-step drug synthesis.

A screenshot of a cell phone

Description automatically generated

**Figure 5**. **|** **Bromhexine Derivative Virtual Library Analysis.** **a)** Synthesis of bromhexine based on retrosynthetic analysis of C–H arylation reactions **b)** Enumeration of 1,000 aryl bromides to compose bromhexine derivative library. **c)** Medicinal chemistry descriptors of bromhexine derivative library. **d)** Morgan fingerprint analysis of bromhexine derivative library.

**Methods**

In the analysis reported herein, the DrugBank dataset and commercial catalogs were filtered to identify compounds with molar mass less than 500 and 200, respectively, and desalted (e.g., removing HCl from amines sold as hydrochloride salts). A Python script (accessible to the reader via an online repository) was written to cross reference targets (such as molecules sampled from DrugBank) with commercially available building blocks. This script accepts a set of user-defined options and requires the location of a commercial catalog and the SMILES of a target molecule to produce a one-step disconnection map between the catalog and target molecules. Building blocks are defined by its substructure SMARTS and a label (e.g., the substructure pattern [B:3](O)[O] describes a boronate and would be labeled as ‘boronate’).

The algorithm is briefly described:

1. Identify target bond in target compound dataset. This target bond is provided via user parameters (e.g., targeting an aryl–aryl C–C bond requires the user to provide the bond substructure SMARTS [c:1]–[c:2] to the script. Store all compounds containing the target bond.
2. Decompose the target bond for all hits, forming two synthons. This is done by recreating the molecule without the targeted bond, leaving virtual R-groups at both points of disconnection. As this is done for each target bond found in the molecule, multiple sets of disconnections can be generated for a single molecule.
3. Enumerate the synthons at the disconnection point to generate building block analogs. The enumerated building block analogs are provided as a list of SMARTS describing the structure of the building block to the script by the user.
4. Cross-reference the enumerated building blocks against the commercial catalog by comparing their unmapped canonicalized smiles. Store the transformation if both synthons can be found as commercial analogs.

**Data Availability**

All data generated from this study, the main reaction targeting script written to generate the data, and the visualization scripts to generate all figures are provided in an online repository.

**Corresponding Author**

\* Tim Cernak – Department of Medicinal Chemistry,

College of Pharmacy, University of Michigan, Ann Arbor,

MI, USA 48104; orcid.org/0000-0001-5407-0643.

Email: [tcernak@med.umich.edu](mailto:tcernak@med.umich.edu)

\* Shannon Stahl – Department of Chemistry,

University of Wisconsin - Madison, Madison,

WI, USA 48104; orcid.org/0000-0002-9000-7665.

Email: [stahl@chem.wisc.edu](mailto:stahl@chem.wisc.edu)

**Reference**

1 Shen, Y., Borowski, J. E., Hardy, M. A., Sarpong, R., Doyle, A. G. & Cernak, T. Automation and computer-assisted planning for chemical synthesis. *Nature Reviews Methods Primers* **1**, 23 (2021). <https://doi.org:10.1038/s43586-021-00022-5>

2 Szymkuć, S., Gajewska, E. P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., Bajczyk, M. & Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie International Edition* **55**, 5904-5937 (2016). <https://doi.org:https://doi.org/10.1002/anie.201506101>

3 Mahjour, B., Shen, Y., Liu, W. & Cernak, T. A map of the amine–carboxylic acid coupling system. *Nature* **580**, 71-75 (2020). <https://doi.org:10.1038/s41586-020-2142-y>

4 Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C. & Wilson, M. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**, D1074-d1082 (2018). <https://doi.org:10.1093/nar/gkx1037>

5 Campeau, L.-C. & Hazari, N. Cross-Coupling and Related Reactions: Connecting Past Success to the Development of New Reactions for the Future. *Organometallics* **38**, 3-35 (2019). <https://doi.org:10.1021/acs.organomet.8b00720>

6 Chen, S.-J., Golden, D. L., Krska, S. W. & Stahl, S. S. Copper-Catalyzed Cross-Coupling of Benzylic C–H Bonds and Azoles with Controlled N-Site Selectivity. *Journal of the American Chemical Society* **143**, 14438-14444 (2021). <https://doi.org:10.1021/jacs.1c07117>

7 Hu, H., Chen, S.-J., Mandal, M., Pratik, S. M., Buss, J. A., Krska, S. W., Cramer, C. J. & Stahl, S. S. Copper-catalysed benzylic C–H coupling with alcohols via radical relay enabled by redox buffering. *Nature Catalysis* **3**, 358-367 (2020). <https://doi.org:10.1038/s41929-020-0425-1>