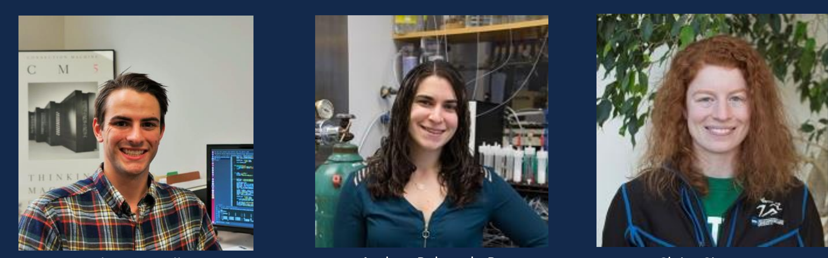
Synthetic Planning and Library Optimization for Automated Cross Coupling Synthesis Platforms

Nathan Russell[1,†], Andrea Palazzolo Ray[2,†] Claire Simons[2], Martin D. Burke[2], Jian Peng[1]
Department of Computer Science[1] & Department of Chemistry[2], University of Illinois at Urbana-Champaign. † Equal Contribution.
Corresponding author: ntrusse2@Illinois.edu



Introduction

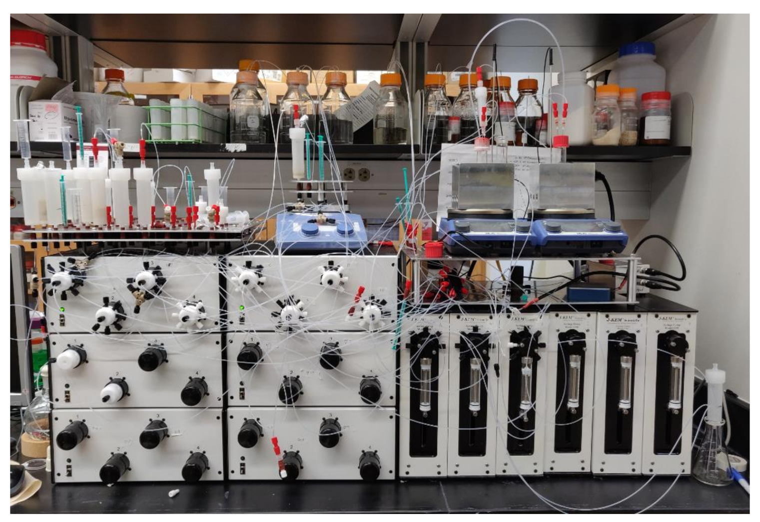
The Goal: Automatable Synthesis of Most Natural Product (NP) Chemical Space via Iterative Cross Coupling (ICC)

Our Contributions:

1. Algorithms for ICC compatible synthesis planning of large target libraries and building block library selection
2. Impact prioritized list of building blocks which enable efficient synthesis of most link-cyclic NP chemical space
3. Characterization of cross coupling reactions and substrate scopes ranked by impact to synthesis of NP chemical space

The Approach:

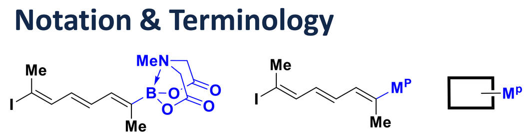
1. Gather a representative set of 100k Link-cyclic Natural Products
2. Given the constraints of iterative cross coupling, generate synthetic plans, allowing all valid blocks and couplings
3. Find the smallest block library which enables coverage of the most NP chemical space using the fewest reaction steps
4. Given block set, identify an "optimal" synthetic plan per target, and record the coupling reactions involved



Automated Cross Coupling Synthesis Platform

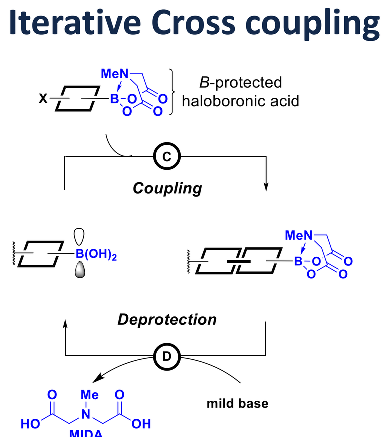
Synthesis Constraints

Notation & Terminology



Termini: reactive group facilitates and is consumed during coupling, possibly protected
Block: molecular fragment with preinstalled termini

Iterative Cross coupling

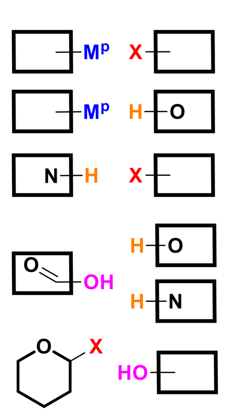


Coupling Reactions Allowed

Cross Couplings: Suzuki-Miyuara, Buchwald-Hartwig, Hiyama-Denmark, Negishi, Sonigashira, Stille, etc.

Heteroatom Acylations

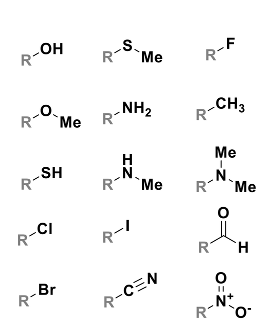
Glycosylations



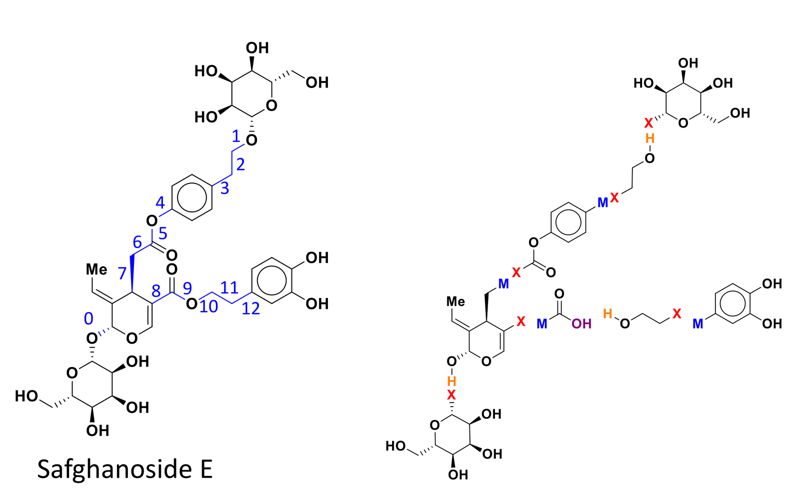
Primary Synthesis Constraints

1. Only Rotatable Bonds can be broken
2. Bonds to R Groups can NOT be broken
3. Blocks must have at least 2 heavy atoms
4. Blocks can have at most 3 Termini*
5. An atom can have at most 2 termini
6. Post Coupling Product, must have at least 1 of the following M^{P1}, M^{P2}, OH, H, or equivalent for Purification turnover unless terminal coupling in plan

R Groups



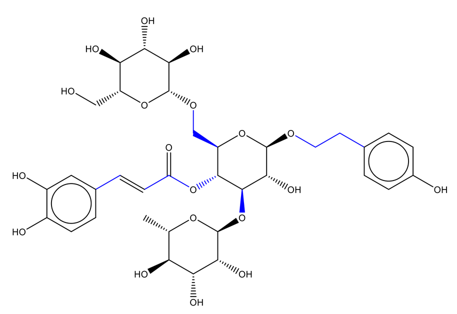
Example Synthetic Plan



Safghanoside E

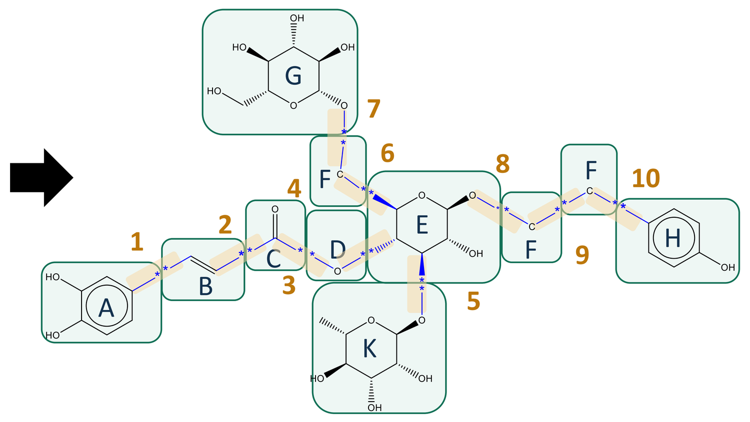
Molecule Representation

1. Target Molecule

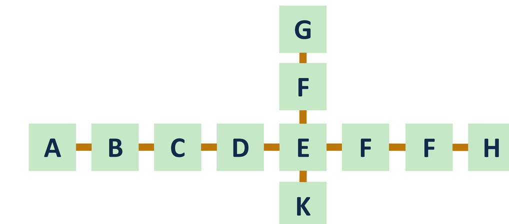


salidroside

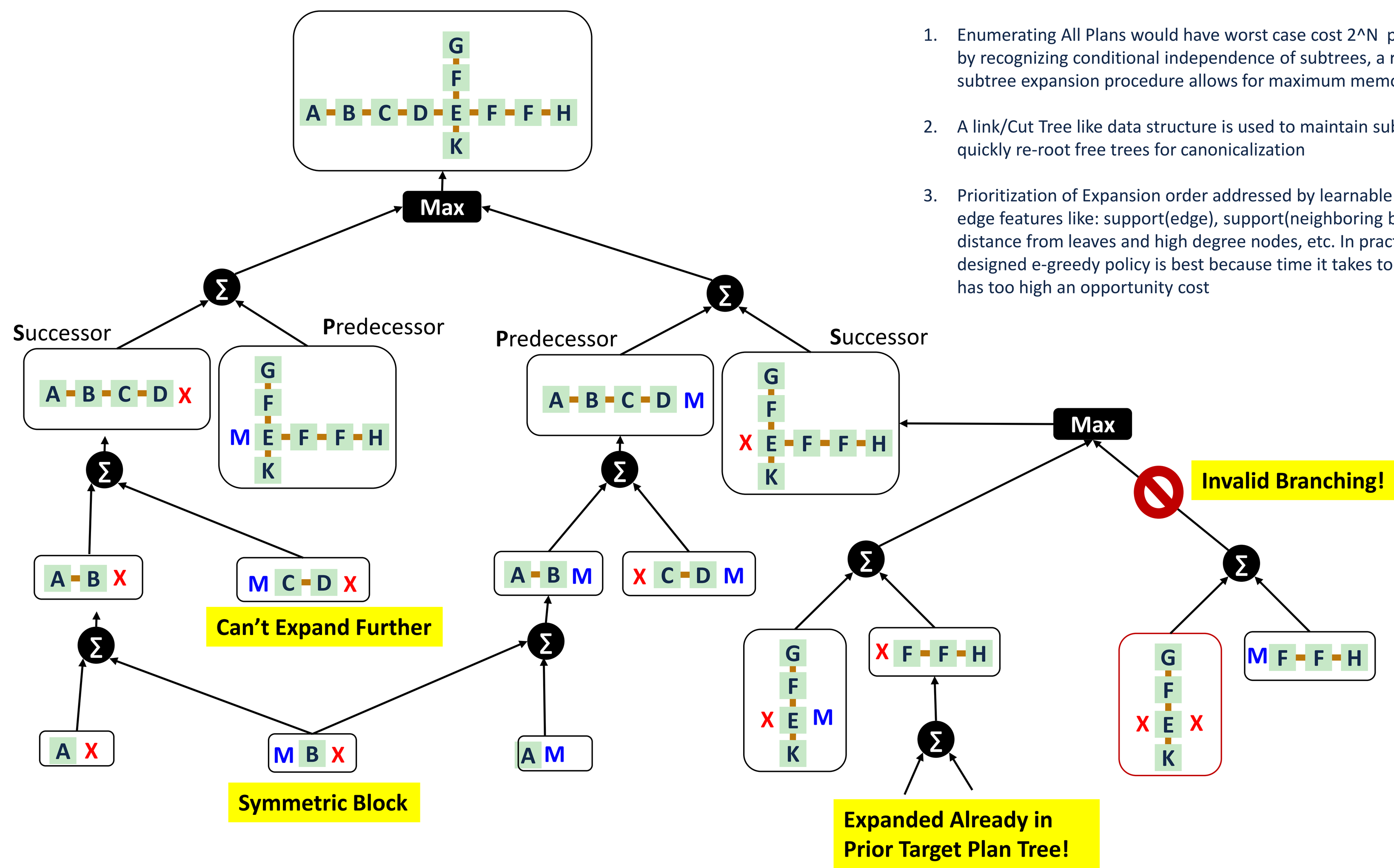
2. Reduction to Free Tree



3.Canonicalization of Subgraphs



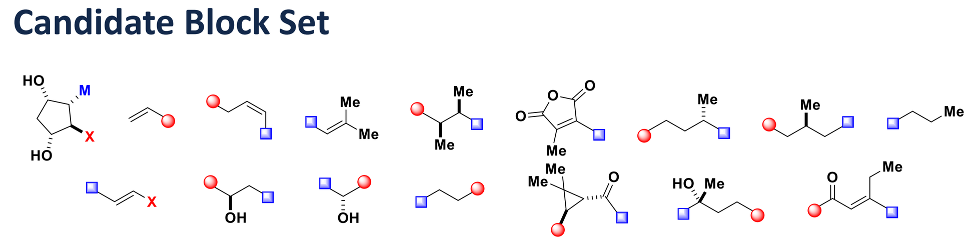
Synthesis Plan Tree Expansion



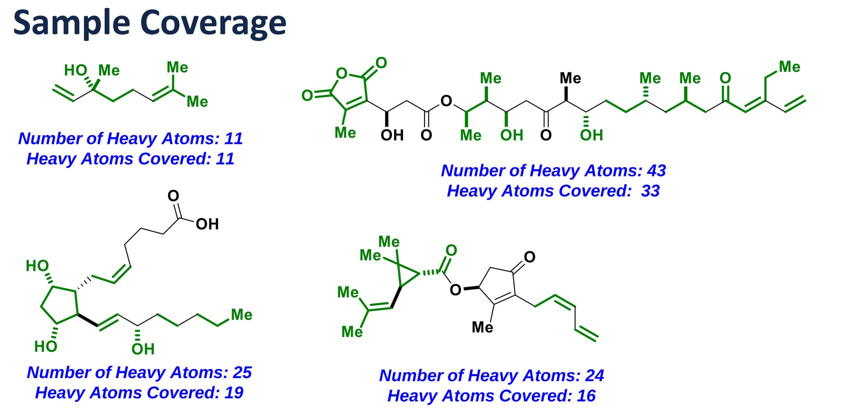
1. Enumerating All Plans would have worst case cost 2^N per target but by recognizing conditional independence of subtrees, a recursive subtree expansion procedure allows for maximum memorization
2. A link/Cut Tree like data structure is used to maintain subtrees and quickly re-root free trees for canonicalization
3. Prioritization of Expansion order addressed by learnable policy over edge features like: support(edge), support(neighboring block), distance from leaves and high degree nodes, etc. In practice, hand designed e-greedy policy is best because time it takes to learn policy has too high an opportunity cost

Optimization / Block + Plan Selection

Candidate Block Set



Sample Coverage



Algorithm: Bidirectional Lazy Greedy Subset Selection for Blocks

```
Sort blocks by descending marginal gain per block, maintain in a MaxHeap
While total_coverage < N%
    For f Forward Steps:
        max_gain = 0
        for block in sorted_block_heap:
            gain = conditional_marginal_gain(block , A)
            max_gain = max(max_gain, gain)
            if max_gain >= UB_gain:
                Add max_block, break
        Add max_block

    For b Backward Steps:
        add block with min loss upon removal from A
```

Optimization Problem

Coverage — *Library Size* — *Tractability*

$$\max_x \sum_{i=1}^{N} z^i \qquad \min_x \sum_{k=1}^{K} x^k \qquad \max_x \sum_{i=1}^{N} \sum_{y_j^i \in SP^i} (y_j^i - 1) nc_j^i$$

s.t.

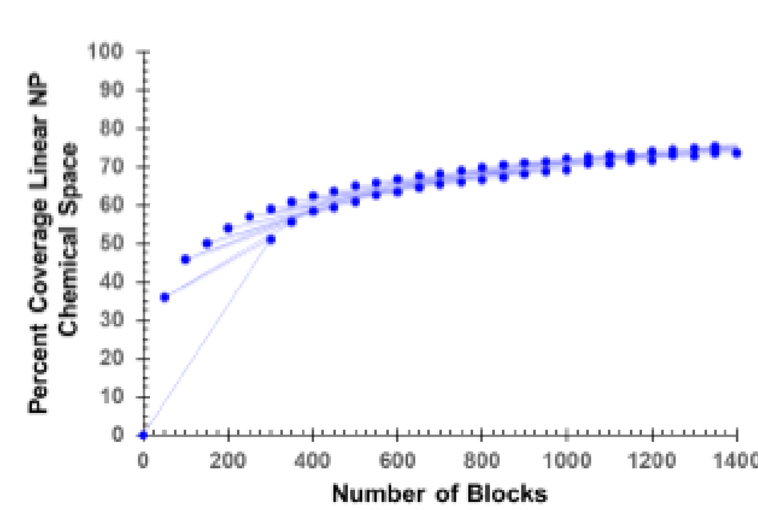
$$z^i \le \vec{x}_j^i \vec{w}_j^{i\top} + M y_j^i \quad \forall\, j \text{ in } SP^i \qquad x^k \in (0,1)\ \forall\, k$$

$$y_j^i \in (0,1)\ \forall\, i,j$$

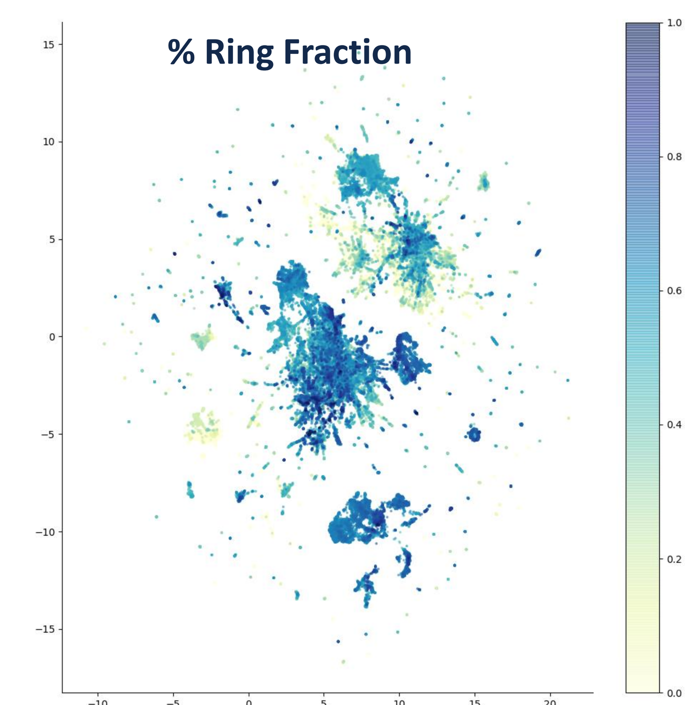
$$\sum_{y_j^i \in SP^i} y_j^i = |SP^i| - 1 \quad \forall\, i$$

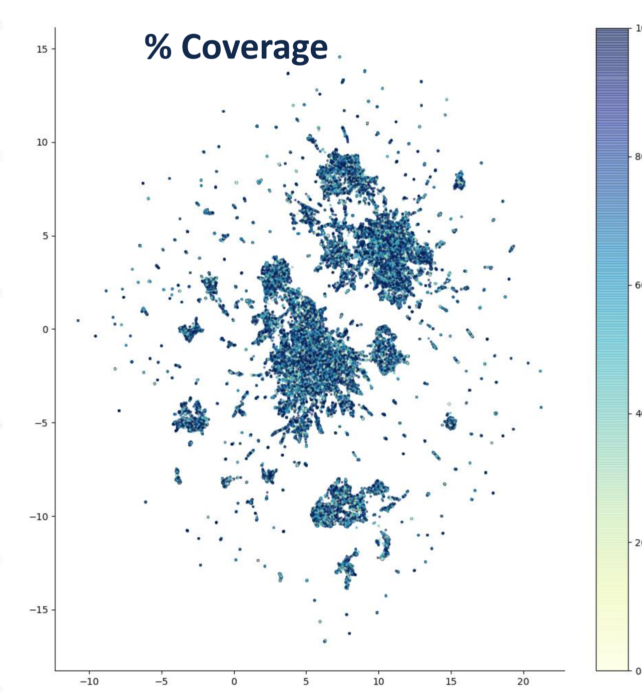
nc_j^i *is the number of couplings for synthetic plan i, j*
z^i *is the maximum covergae achivable for natural product i*
$\vec{x}_j^i$ *is the vector of block decision variables for synthetic plan i, j*
$\vec{w}_j^i$ *is the vector of heavy atom count constants for synthetic plan i, j*
M *is a Constant* $\ge \max(\vec{x}_j^i \vec{w}_j^{i\top}) - \min(\vec{x}_j^i \vec{w}_j^{i\top})\ \forall\, i,j$
$N \ge 10^5$: # of Natural Products, indexed by i
$K \ge 10^7$: # of Block Candidates, indexed by k
$SP^i \propto 2^{n\ breakable}$: *Set of candidate synthetic plans*, plans indexed by j.

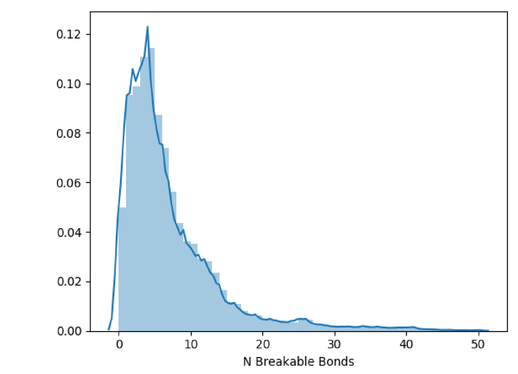
Optimization Curve



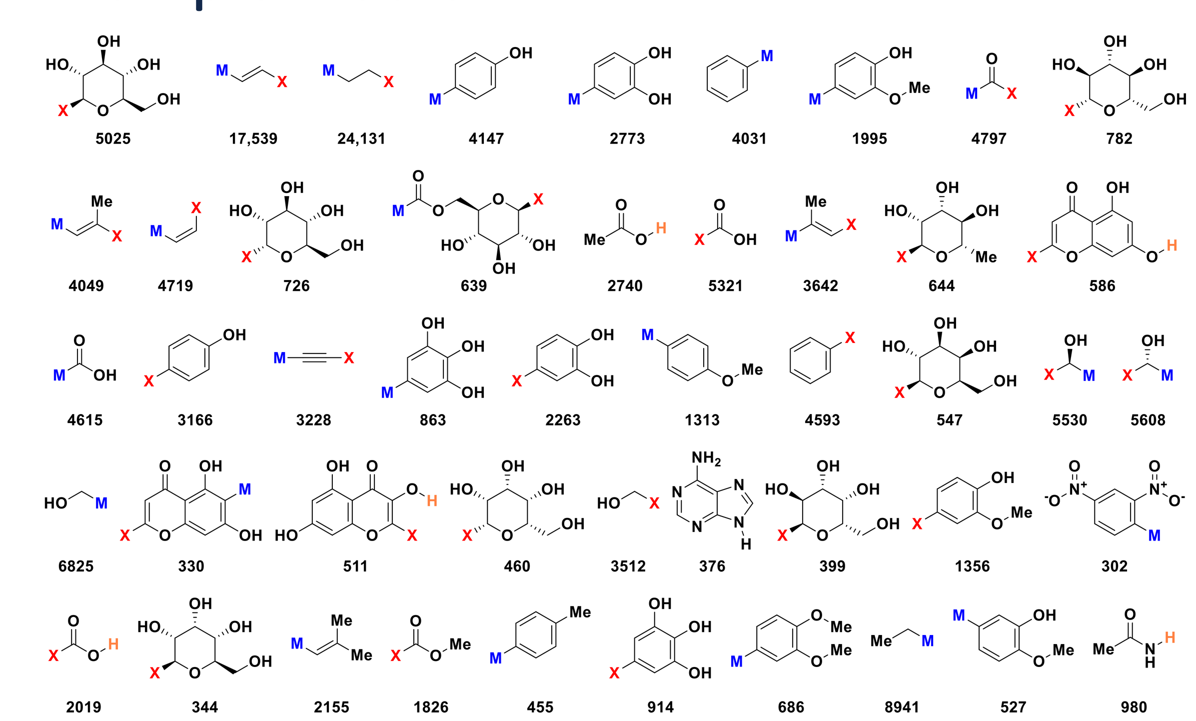
Natural Product Space

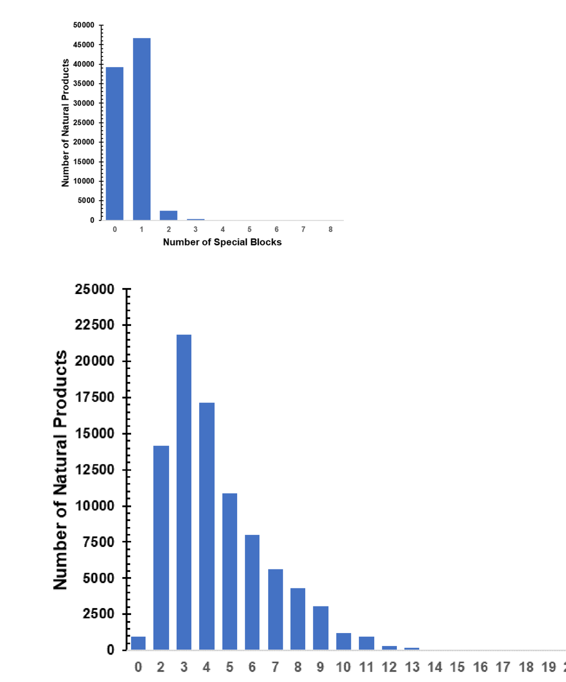




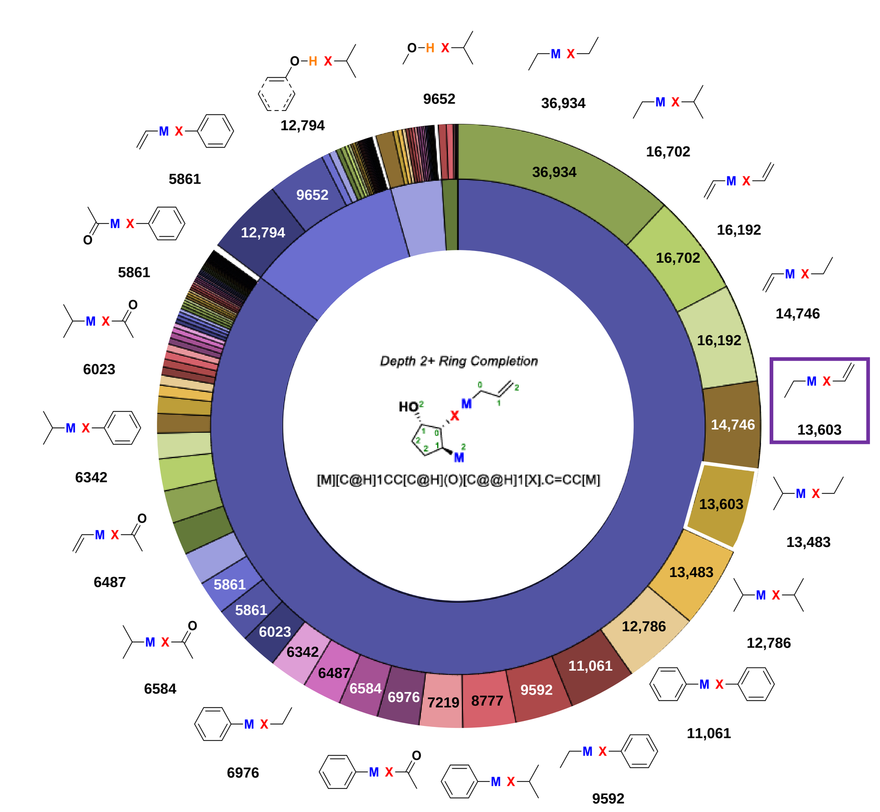


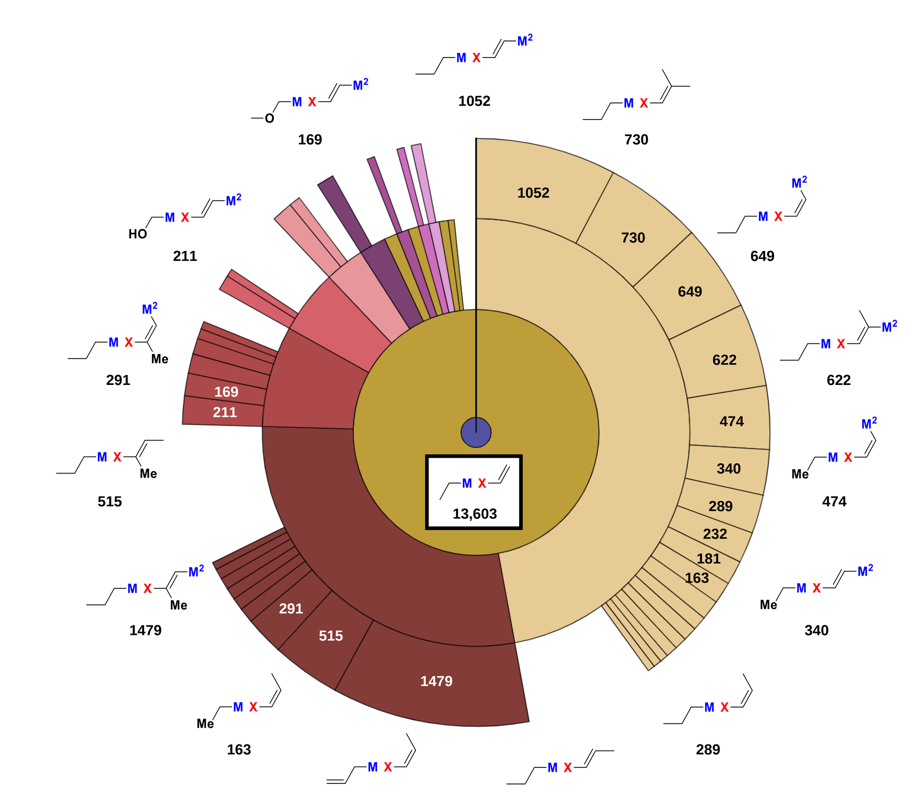
Example Blocks

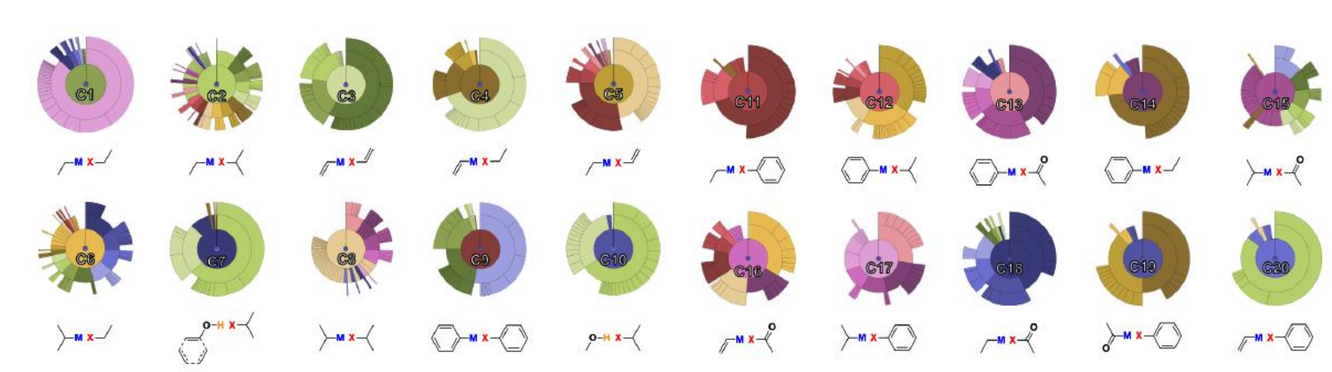


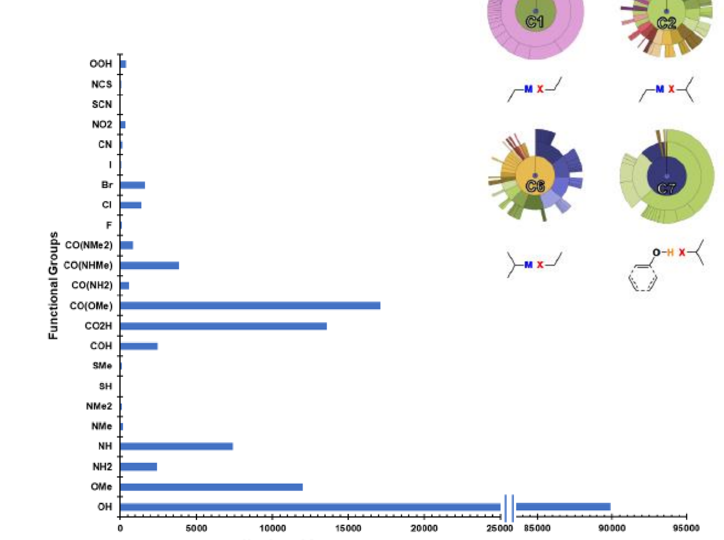


Prioritized Cross Coupling Substrate Scopes









Acknowledgements