Supporting Information

RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application

Connor W. Coley, William H. Green*, and Klavs F. Jensen*

Department of Chemical Engineering, MIT, Cambridge, MA 02139

E-mail: whgreen@mit.edu; kfjensen@mit.edu

Code

All code can be found at `https://github.com/connorcoley/rdchiral` in addition to Jupyter notebooks containing all of the examples included in this manuscript.

Additional Results

Special Functional Group List

Table S1: List of special functional groups currently used for template extraction to ensure important motifs are included in the template. Atom IDs refers to the indices of the atoms in the SMARTS pattern that will trigger inclusion of that motif. For example, only the atom matching the wildcard in `C#C-[*]` triggers inclusion of that motif; the alkynyl carbons will not trigger inclusion of their neighbors. This list is not meant to be exhaustive or definitive and is easily modified in the code.

SMARTS	Atom IDs	Description
`[OH0,SH0]=C[O,Cl,I,Br,F]`	All	carboxylic acid / halogen
`[OH0,SH0]=CN`	All	amide/sulfamide
`S(O)(O)[Cl]`	All	sulfonyl chloride
`B(O)O`	All	boronic acid/ester
`[Si](C)(C)C`	0	trialkyl silane
`[Si](OC)(OC)(OC)`	0	trialkoxy silane, default to methyl
`[N;H0;$(N-[#6]);D2]-,=[N;D2]-,=[N;D1]`	All	azide
`O=C1N([Br,I,F,Cl])C(=O)CC1`	All	NBS-like reagent
`Cc1ccc(S(=O)(=O)O)cc1`	All	Tosyl
`CC(C)(C)OC(=O)[N]`	7	NBoc
`[CH3][CH0]([CH3])([CH3])O`	4	trimethylmethoxy
`[C,N]=[C,N]`	All	alkene/imine
`[C,N]#[C,N]`	All	alkyne/nitrile
`C=C-[*]`	2	adj to alkene
`C#C-[*]`	2	adj to alkyne
`O=C-[*]`	2	adj to carbonyl
`O=C([CH3])-[*]`	3	adj to methyl ketone
`O=C([O,N])-[*]`	3	adj to carboxylic acid/amide/ester
`ClS(Cl)=O`	All	thionyl chloride
`[Mg,Li,Zn,Sn][Br,Cl,I,F]`	All	grinard/metal (non-disassociated)
`S(O)(O)`	All	SO2 group
`N~N`	All	incl. diazo
`[!#6;R]@[#6;R]`	1	adj to heteroatom in ring
`[a!c]:a:a`	2	two-steps from aromatic heteroatom
`[B,C](F)(F)F`	0	CF3, BF3 should have the F3 included

Template Extraction

The ring closing reaction in Figure S1A, as defined by the reaction SMILES, requires the trans cyclobutane in the precursor. The ketone in Figure S1B is prepared from the chiral alcohol, so the corresponding template specifies a chiral alcohol precursor. Figure S1C defines a reaction

between an alkyl iodide and the β position of an α,β-unsatured ketone. Figure S1D defines the preparation of a bromohydrin from a cis epoxide. And, finally, Figure S1E describes the stereoinversion of a chiral mesylate with an azide via an S_N2 reaction.

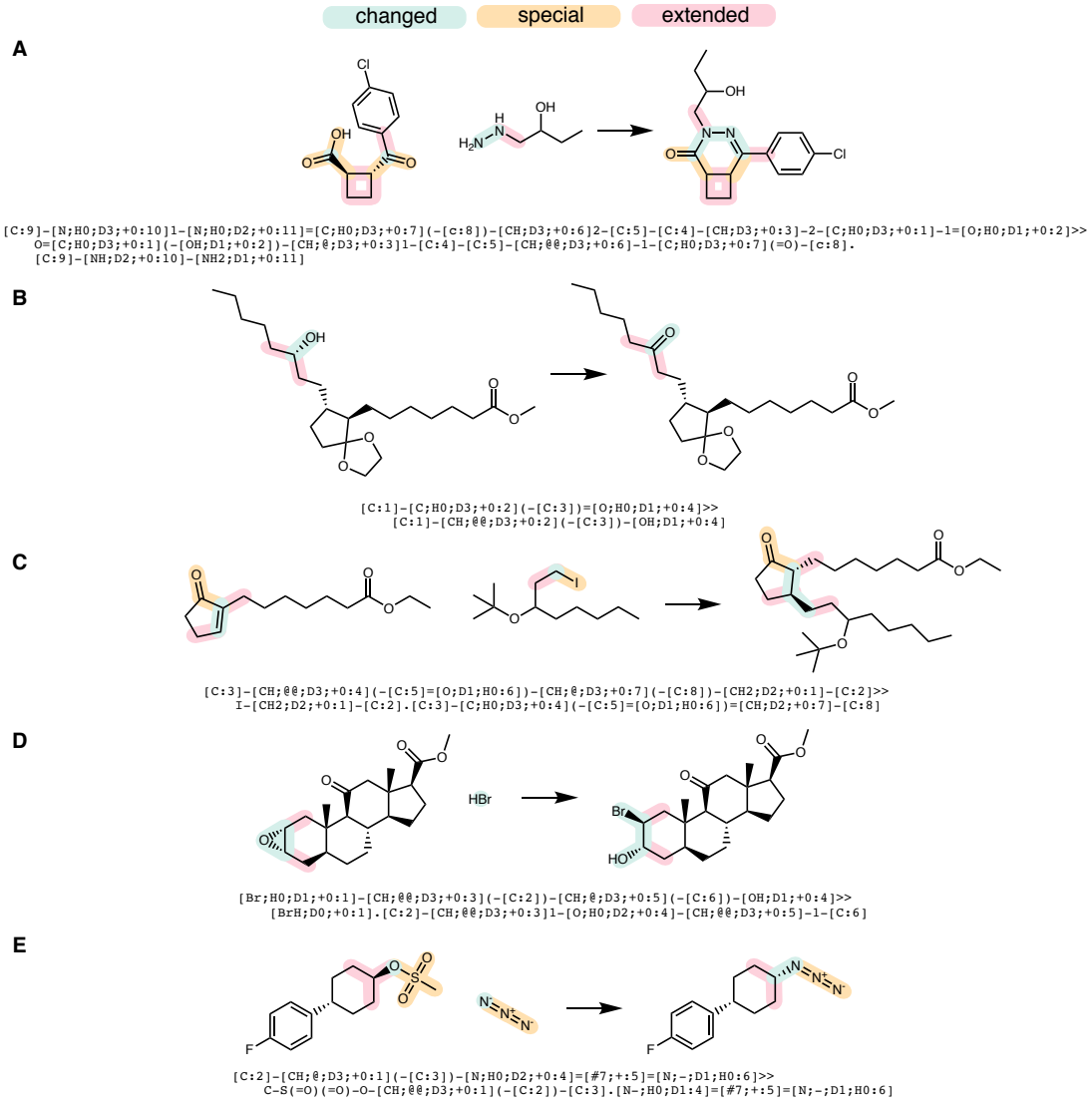

Figure S1: Additional examples of reactions from the USPTO and the retrosynthetic templates we extract from them, where stereochemistry must be considered. Atom mapping is omitted for brevity, but is unambiguous in all cases shown. Spectator molecules are also omitted, as they do not contribute heavy atoms to the product and are not included in the resulting reaction template.

Data Quality Issues

The workflow relies on the availability of atom-mapped reaction SMILES strings. There are cases where the quality of examples can lead to nonsensical reaction templates. Two such examples are shown in Figure S2. The first, shown in Figure S2A, is a result of poor atom mapping. The reaction is a substitution of an alkyl chloride using sodium cyanide to prepare the alkyl cyanide. However, the solvent dimethylsulfoxide (DMSO) is labeled as contributing to the heterocycle in the product molecule. To the algorithm, all mapped atoms aside from atom 3 appear to undergo a change in connectivity, which causes the resulting template to include the entire product molecule.

The second example, shown in Figure S2B, appears to be an erroneous reaction example entirely. It is likely that the unlabeled molecule was meant to be the reactant and the labeled reactant was meant to be the product of an oxidation. However, given the entry and its atom mapping, the apparent reaction is a shortening of the propyl side chain to ethyl. The corresponding template describes a reaction where any alkyl chain of at least two carbons can be synthesized from the same compound with that chain extended by a single carbon.

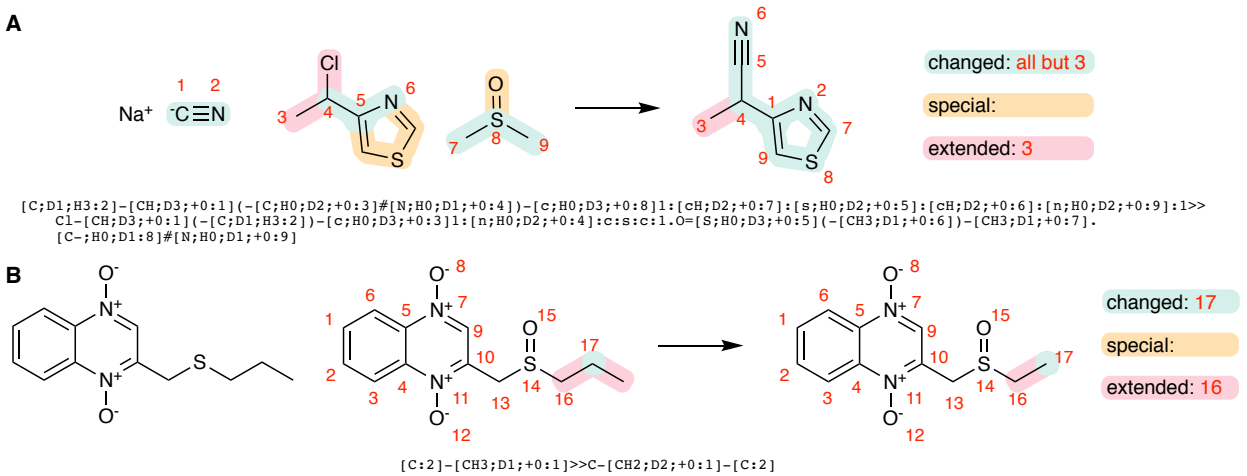

Figure S2: Reactions where the extracted template SMARTS is not chemically meaningful. (A) A case of poor atom mapping, where what should be a simple substitution reaction results in a template that encompasses the entire product structure. (B) A case of poor data quality, where the resulting template suggests that any alkyl chain of length $n \geq 2$ can be prepared from an alkyl chain of length $n + 1$.