Computer-Assisted Retrosynthesis Based on Molecular Similarity

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

2021.5.14

Outlines

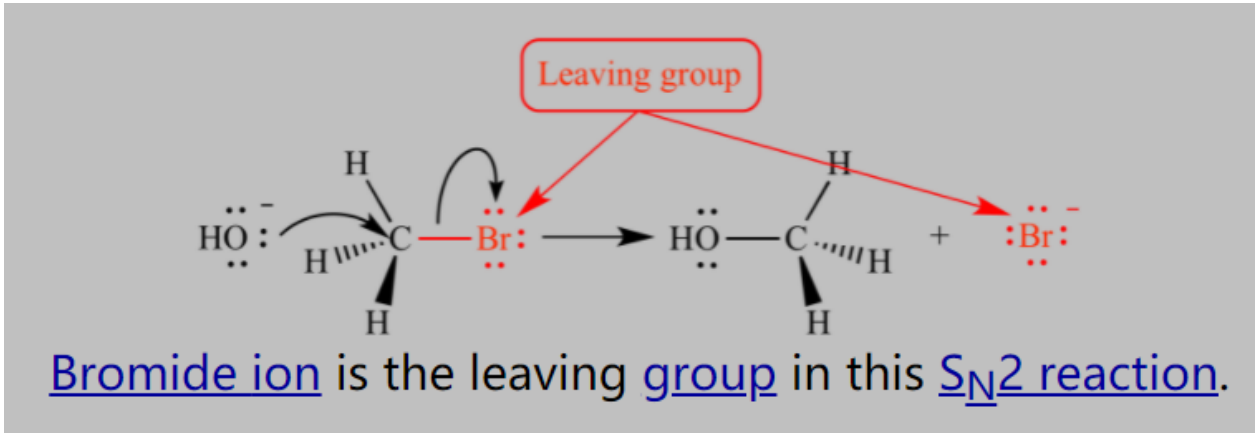
Retrosynthesis

To synthesize a target chemical compound, it is necessary to identify a series of suitable reaction steps beginning from available starting materials.
This analysis ——starting from the target compound and working backward was later formalized as retrosynthesis by E. J. Corey.

Limitations of Existing Template-based Approaches

Most automated retrosynthesis programs have relied on encoding reaction templates, or generalized subgraph matching rules.

- abundance of distinct leaving groups for equivalent reaction sites in a single template
- computationally expensive due to the cost of solving the subgraph isomorphism problem



Contribution

This paper proposes and validates a similarity-based approach.

- In the new approach, strategic disconnections are performed based solely on analogy to known reaction precedents. So it's purely deterministic and does not require tuning or training of any model parameters.
- It fully specifies leaving groups when extracting and applying templates from precedent reactions.

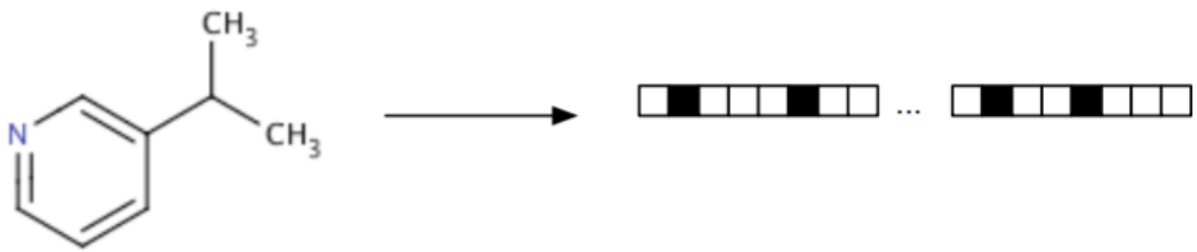
SMILES and SMARTS

- The SMILES notation allows one to represent a 2D chemical drawing as a string, (e.g. ”C1CCCCC1” for cyclohexane).
- SMARTS is an extension of SMILES that allows one to specify chemical patterns with wildcards for atoms or bonds, e.g. ”[C,N,O]?” .
- SMARTS and SMILES are intended for fundamentally different things:
  - SMILES is to represent particular compounds.
  - SMARTS is more general. It represents a query against a range of possible molecules, or an abstract description of a set of possible molecules.
  - Example:
    ’CC’ as a SMILES string depicts a single compound: ethane(乙烷).
    As a SMARTS query, CC will match ethane(乙烷), but will also match propane(丙烷), acetic acid(乙酸), cyclohexane(环己烷), vancomycin(万古霉素), etc.

Molecular Fingerprint

Molecular fingerprints are a way of encoding the structure of a molecule.

- Idea: Apply a kernel to a molecule to generate a bit vector or count vector (less frequent).
- The most common type of fingerprint is a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule.
- Typical kernels extract features of the molecule, hash them, and use the hash to determine bits that should be set.
- Lots of experience shows that the best fingerprint depends strongly on the data set. So there are many different fingerprints available. (e.g. Morgan/Circular fingerprints)
- Comparing fingerprints allows you to determine the similarity between two molecules, to find matches to a query substructure, etc.



Similarity Calculation

Quantifying molecular similarity on the basis of two- dimensional (2D) structure generally requires a fingerprinting technique (to represent a molecule as a vector) and a similarity metric (to compare the two vectors of two molecules)

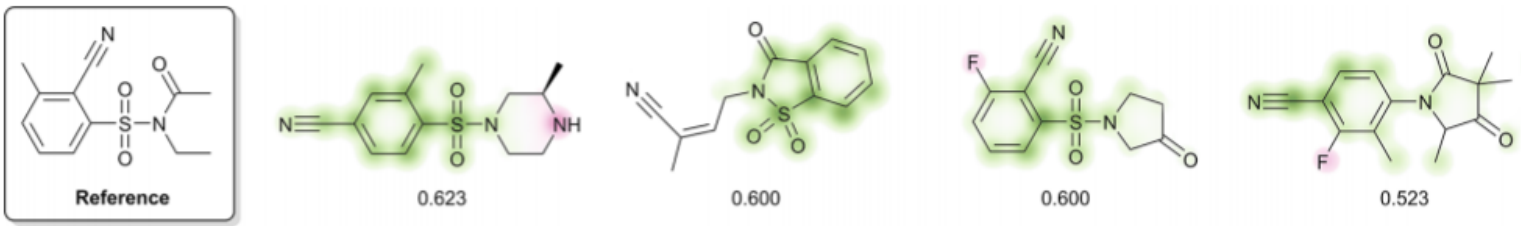


Figure 3. Example similarity score calculation using *Morgan2Feat* fingerprints and the *Tanimoto* metric. Colors indicate atom-level contributions to the overall similarity (green: increases similarity score, red: decreases similarity score, uncolored: has no effect).

Similarity Metrics

Given 2 fingerprint vectors $\mathbf{x}$ and $\mathbf{y}$,

$$\text{Dice}(\mathbf{x}, \mathbf{y}) = \frac{2 \sum x_i y_i}{\sum x_i^2 + \sum y_i^2}$$

$$\text{Tanimoto}(\mathbf{x}, \mathbf{y}) = \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i v_i}$$

$$\text{Tversky}(\mathbf{x}, \mathbf{y}; \alpha, \beta) = \frac{\sum x_i y_i}{\alpha \sum x_i^2 + \beta \sum y_i^2 - \sum x y_i}$$

Overview

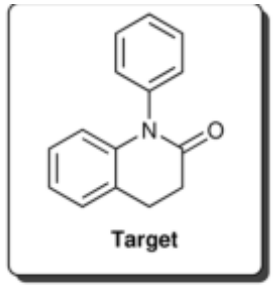
Fact: Similar products tend to be produced by similar reactions.
How have similar molecules been synthesized?

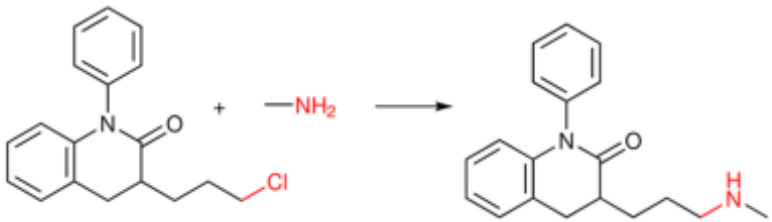
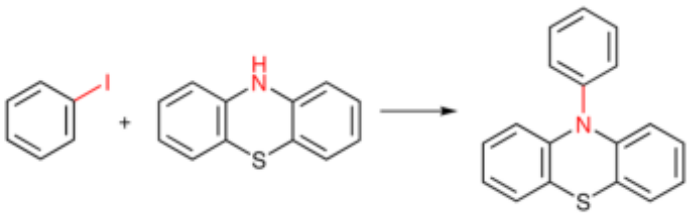
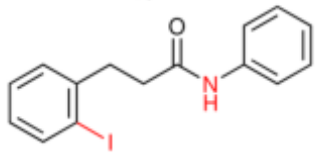
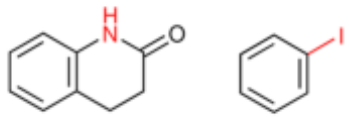
- If a route to the molecule has been previously published, it may be appropriate to use that route without modification.
- If it is a novel compound, then one might look at routes to other compounds with similar structural motifs and determine whether that synthetic strategy is applicable.

Approach

1. Retrieve reaction precedents from the knowledge base based on product similarity, s_{prod}, scored between 0 and 1.
2. Extract a highly local transform containing fully specified leaving groups from each precedent reaction and apply to the target compound.
3. Calculate the similarity of candidate precursors to that precedent's reactants, s_{reac}, between 0 and 1.
4. Rank the resulting candidates by the overall similarity score as calculated by
$$s = s_{prod} \dot{s}_{reac}.$$
5. Evaluating the performance of the approach by compare each of the candidate precursor's isomeric SMILES representation to the target compound's known reactant SMILES string.

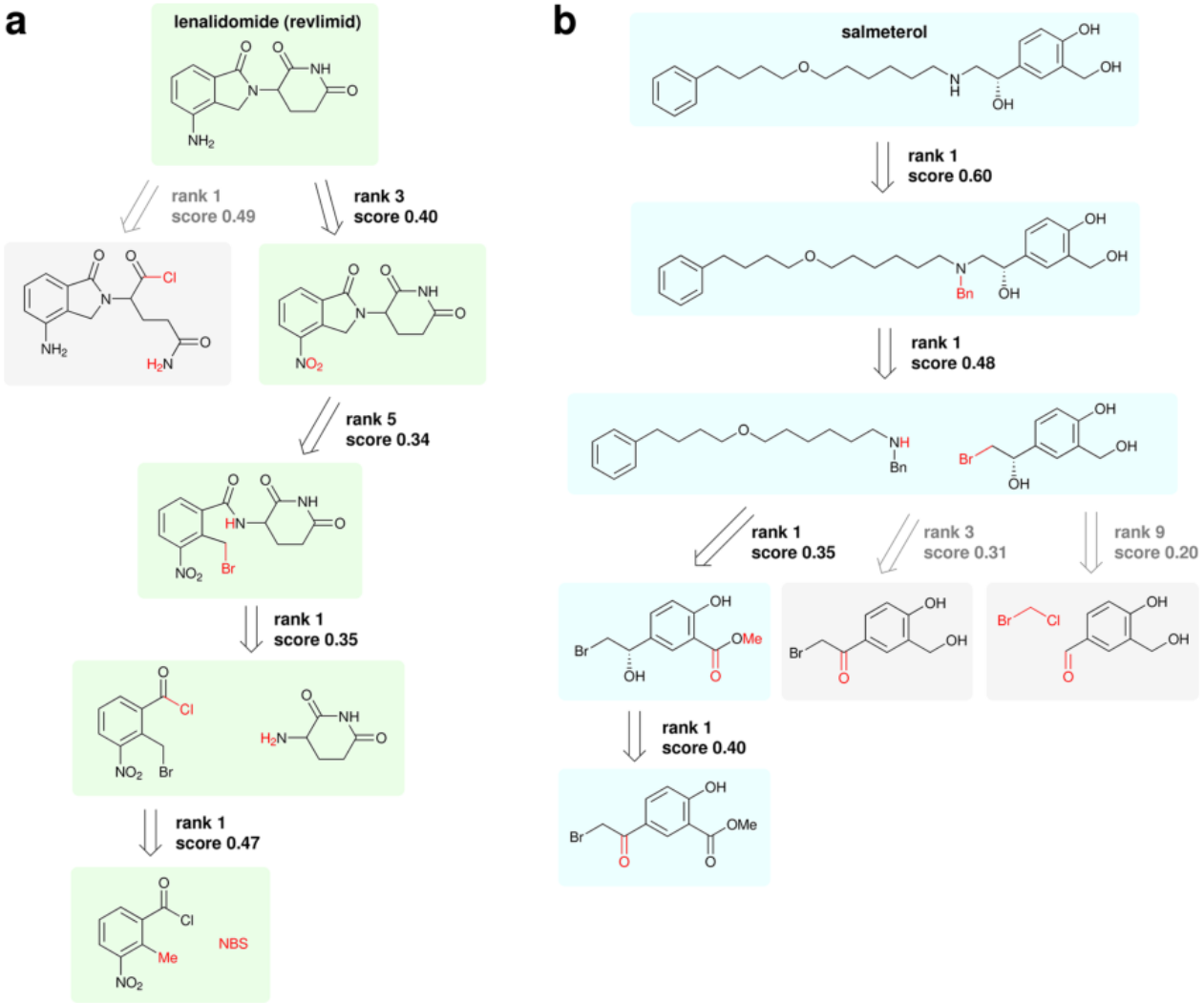
Approach



Reaction Precedent (sorted by product similarity)	Reaction Site	Precursor Sugestion(s)	Final Rank *(score)*
Precedent 1, *product similarity: 0.562* 	[NH1]—[CH2]	*Reaction site [NH1]-[CH2] does not match this target, so no precursors are generated from this precedent*	
Precedent 2 does not have a reaction site applicable to this target			
Precedent 3, *product similarity: 0.467* 	[NH0]—[cH0]	**Precursor 3a**, *precursor similarity: 0.326*  **Precursor 3b**, *precursor similarity: 0.481* 	3 *0.152* 1 *0.224*
Precedents 4-10 do not have a reaction site applicable to this target			

Applications

- One-step evaluation
- Multi-step planning



Limitations of the Approach

The similarity-based approach is meant to apply existing reaction knowledge to novel substrates. It's an empirical, data-driven approach to automated retrosynthesis.
The retrosynthetic suggestions do not offer major insights beyond the types of reactions in the knowledge base.

Conclusions

- Computer-assisted retrosynthesis typically involves some high-level strategy to help guide the search toward simpler, buyable chemicals (e.g., favoring smaller precursors).
- Information about reagents(试剂), catalysts(催化剂), solvents(溶剂), and temperatures of precedent reactions should be considered to enrich the retrosynthesis suggestions.