ML-driven pathway design and validation in automated synthesis

Connor W. Coley, Hanyu Gao, Thomas Struble, William H. Green, Klavs F. Jensen
Wengong Jin, Regina Barzilay, Tommi S. Jaakkola

Department of Chemical Engineering, MIT
Computer Science and Artificial Intelligence Laboratory, MIT

Automation gap in chemical synthesis

Advances in laboratory automation promise to decrease the manual effort of synthesis, but determining *how* to synthesize a compound still requires expert chemist knowledge

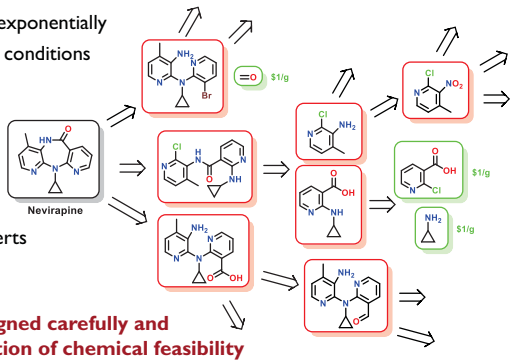
Hwang® and Coley® et al., *Chem. Comm.* **2017**

Godfrey, Masquelin, and Hemmerle, *Drug Discov. Today* **2013**

Computer-aided synthesis planning (CASP)

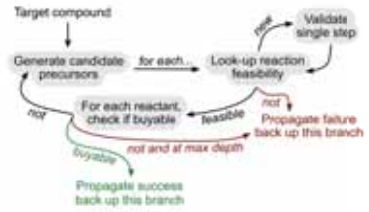
Pathways can be generated through recursive *retrosynthetic* expansion, but…

- The search space grows exponentially
- Suggestions must include conditions in addition to structures
- Disconnection rules are imperfect and may not be chemically feasible
- Detecting and reducing these false positives is challenging, even for experts
- No system has achieved wide-spread adoption



Pathways must be designed carefully and with explicit consideration of chemical feasibility

Extensible framework for pathway design



Key elements include:

- Candidate precursor generation with awareness of chirality
- Reaction condition recommendations
- Feasibility assessment through explicit prediction of reaction outcomes

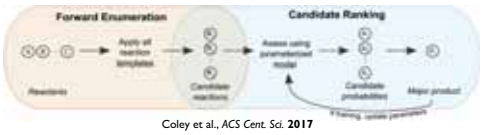
Coley et al., *Acc. Chem. Res.* **2018**

Validating via prediction of reaction outcomes

Retrosynthetic suggestions can be evaluated by solving the reverse problem:

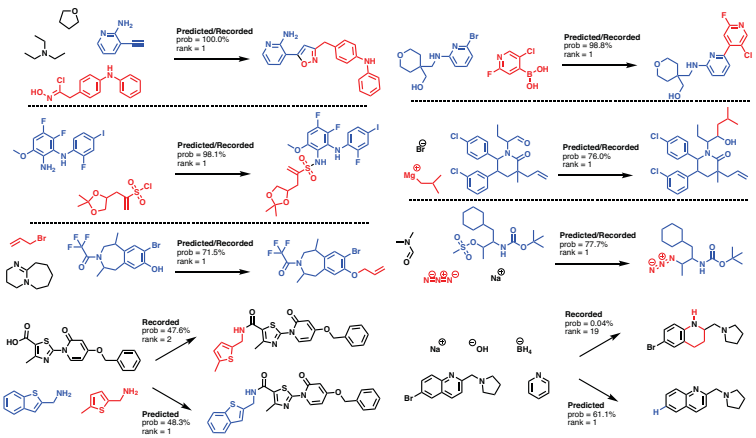
Given a set of reactants (and conditions), what is the major product?

[v1] Template-based, hybrid method



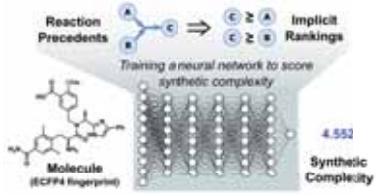
Coley et al., *ACS Cent. Sci.* **2017**

- Reaction templates define scope of possible outcomes
- Neural network model ranks candidates



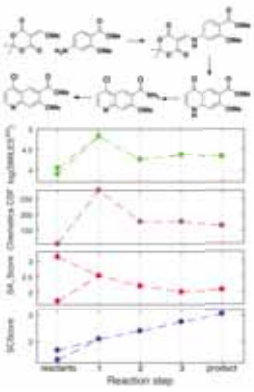
Guiding a retrosynthetic search: SCScore

To overcome combinatorial explosion, the search is focused along promising directions likely to lead back to purchasable compounds



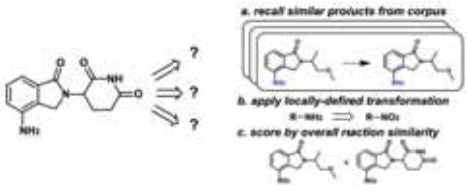
Published reaction data can be used to learn a synthetic complexity score to quantify if progress is being made in a synthetic step

Coley et al., *J. Chem. Inf. Model.* **2018**



Making focused retrosynthetic suggestions

An automated procedure can codify the workflow a beginner chemist might undertake when synthesizing a novel molecule: *look for similar compounds*



Model	top-n accuracy (%), n = 1	3	5	10	20	50
Liu et al. baseline	35.4	52.3	59.1	65.1	68.6	69.5
Liu et al. augmented	37.4	52.4	57.0	61.7	65.9	70.7
Similarity (this work)	**52.9**	**73.8**	**81.2**	**88.1**	**91.8**	**92.9**

- Disconnections proposed based on analogy
- Rare transforms need not be excluded
- Extends to multi-step planning
- Outperforms previous neural translation study

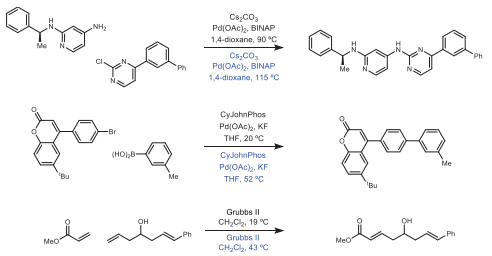
Coley et al., *ACS Cent. Sci.* **2017**

Recommending reaction conditions

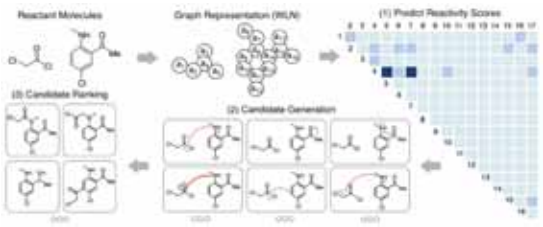
A recommender system suggests conditions for novel reactions

- Compatibility implicitly considered
- Close agreement to literature on test set

Gao, Struble, Coley et al., *submitted*



[v2] Template-free, fully-learned method



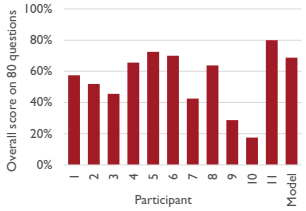
Jin, Coley et al., *NIPS* **2017**

- Template-free method overcomes bottleneck of template application
- Achieves higher accuracy and expert-level performance

USPTO-15K				
Method	Cov.	P@1	P@3	P@5
Coley et al.	100.0	72.1	86.6	90.7
WLN	90.1	74.9	84.6	86.3
WLDN	90.1	**76.7**	**85.6**	**86.8**
USPTO				
Method	|θ|	P@1	P@3	P@5
WLDN	3.2M	**79.6**	**87.7**	**89.2**
WLDN (*)	3.2M	83.9	93.2	95.2

Quantitative performance. Prediction at ranks 1, 3, 5. (*) denotes that the true product was added if not covered during previous stage

Human benchmarking



Research consortium
mlpds.mit.edu

Software demonstrations
askcos.mit.edu

Contact
cwcoley.com