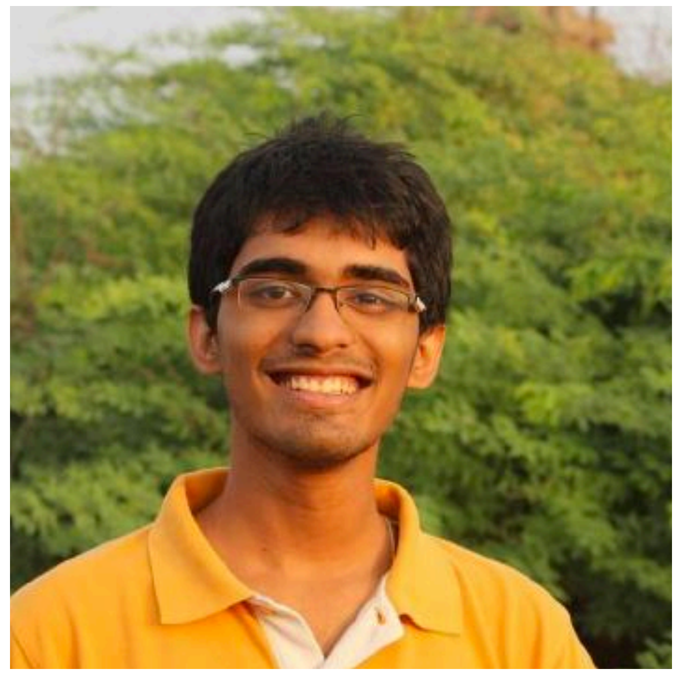
Learning Graph Models for Retrosynthesis Prediction



Vignesh Ram Somnath

Charlotte Bunne

Connor Coley

Andreas Krause

Regina Barzilay

MIT and ETH zürich

Outline

Problem Introduction

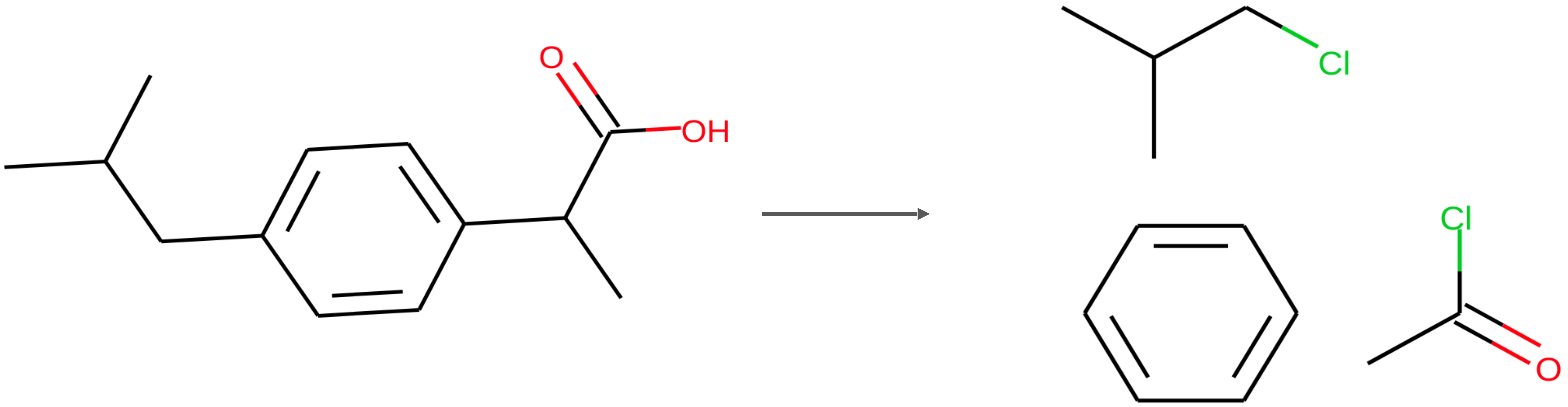
Prior Work

Model Formulation

Experiments and Conclusion

Retrosynthesis

Given a target molecule, predict precursors that can be used to design it

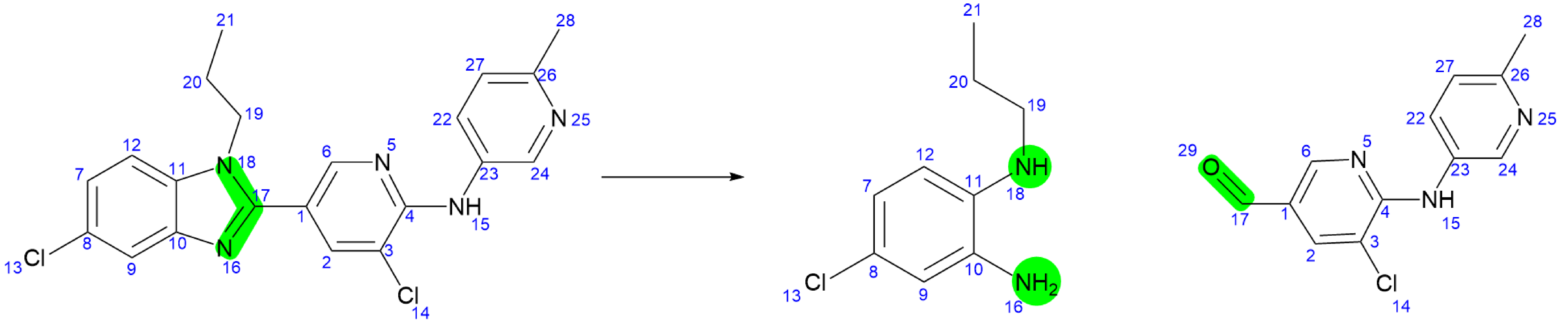


Ibuprofen

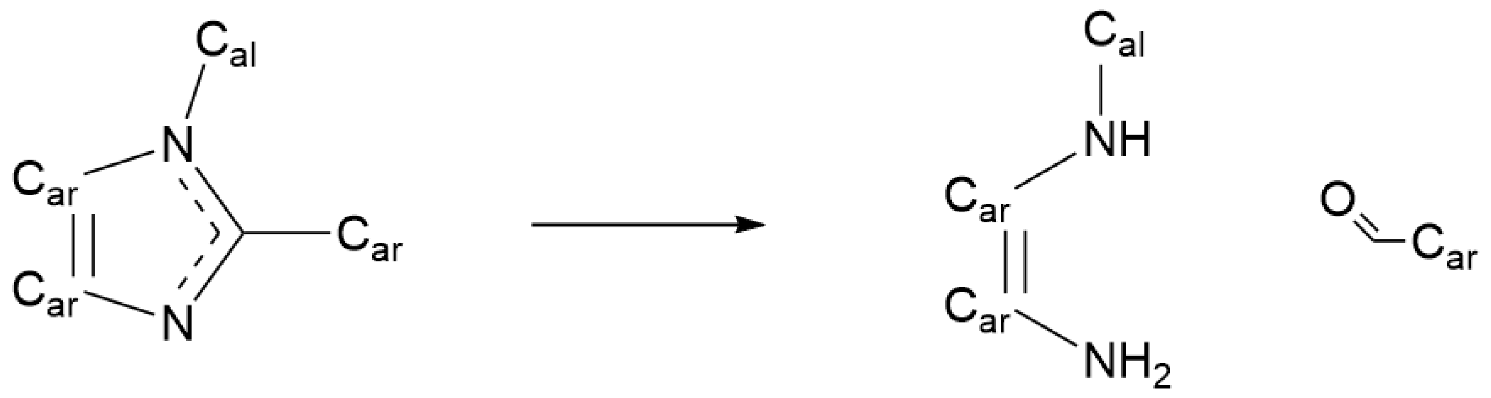
Ibuprofen precursors

Prior Work: Template-Based

Example Reaction



Corresponding Template

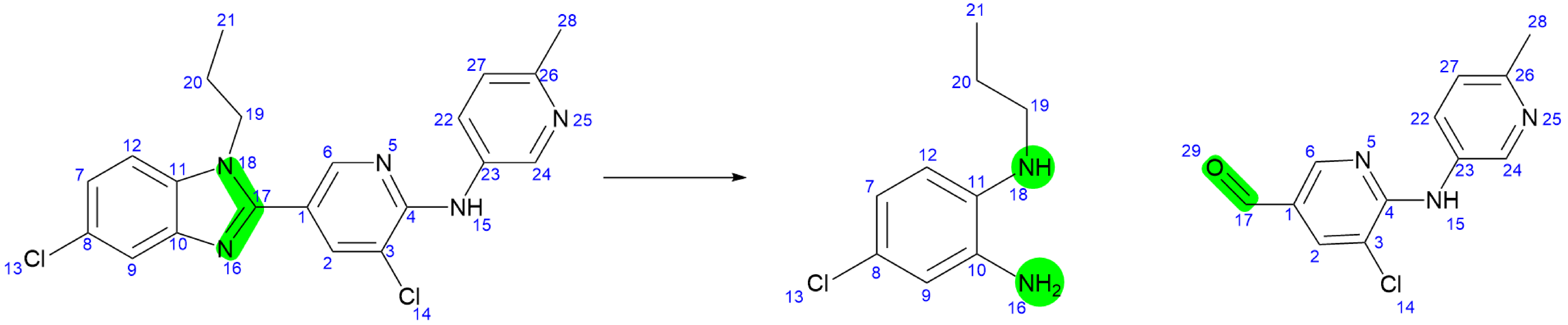


C$_{ar}$ = Aromatic carbon
C$_{al}$ = Aliphatic carbon

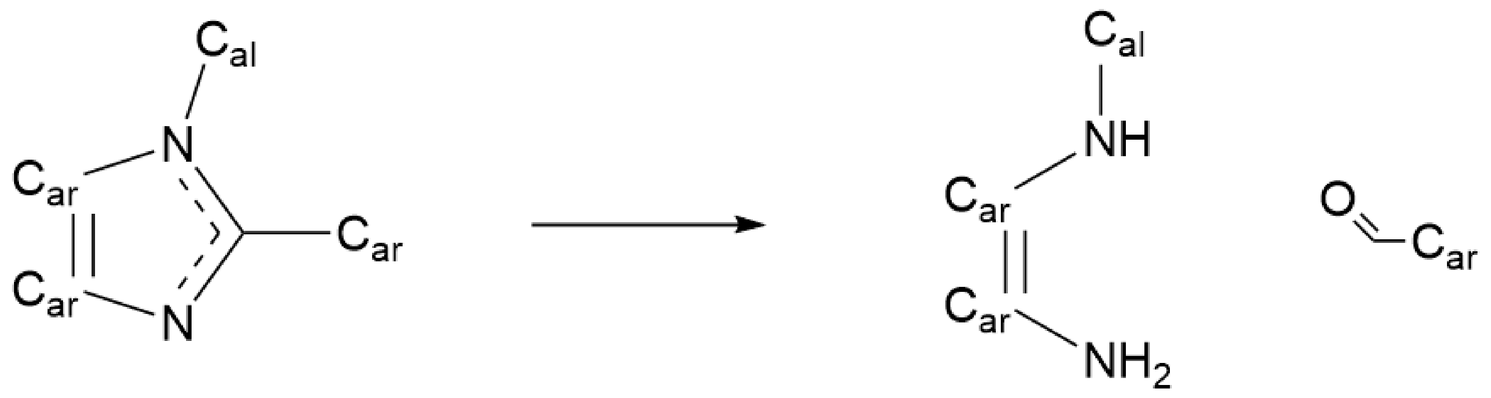
Coley et al. (2017), Segler et al. (2017), Dai et al. (2019)

Prior Work: Template-Based

Example Reaction



Corresponding Template



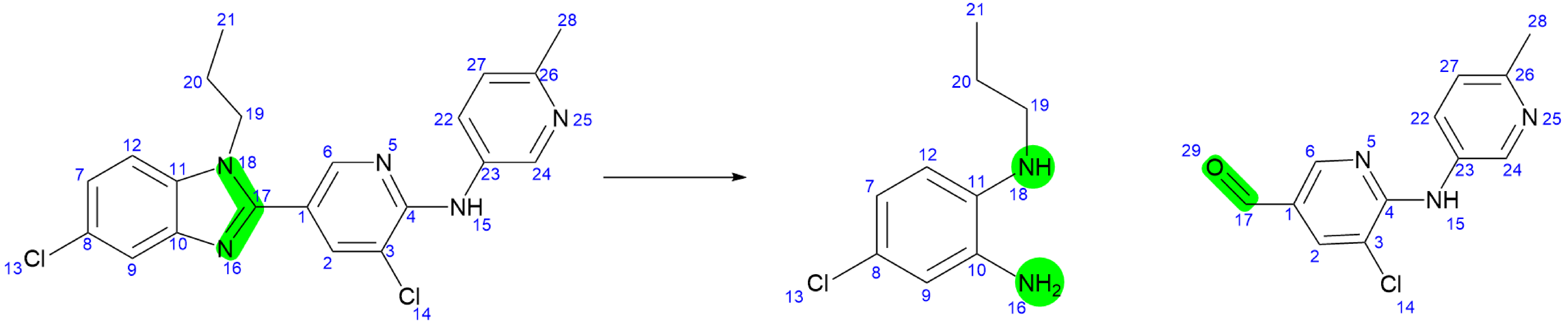
C_{ar} = Aromatic carbon
C_{al} = Aliphatic carbon

Coley et al. (2017), Segler et al. (2017), Dai et al. (2019)

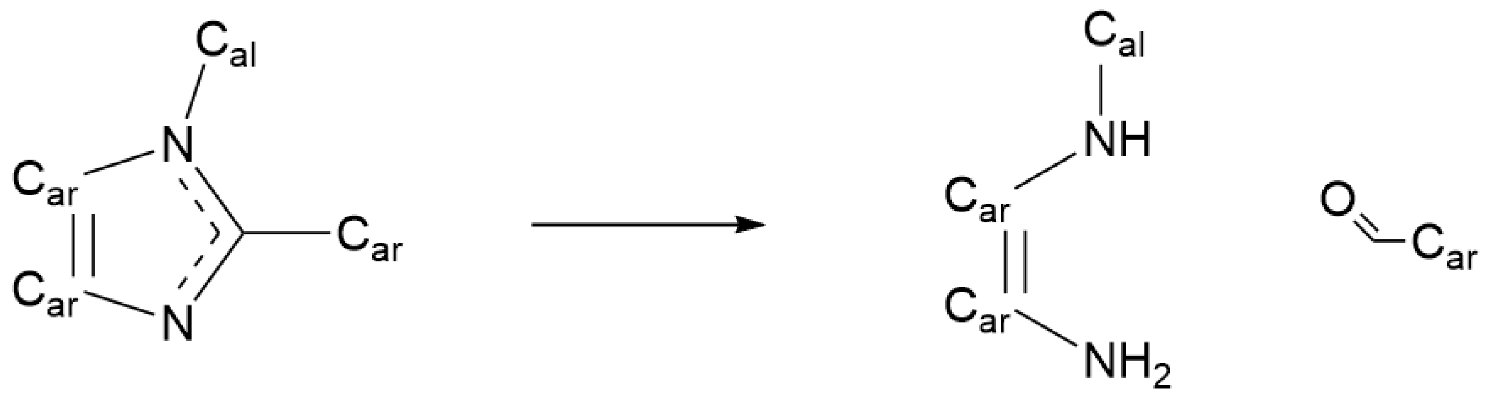
- Coverage vs scalability tradeoff
- Relevance: Rules for a given molecule

Prior Work: Template-Based

Example Reaction



Corresponding Template



C_{ar} = Aromatic carbon

C_{al} = Aliphatic carbon

Coley et al. (2017), Segler et al. (2017), Dai et al. (2019)

Advantages:

- Interpretable - Knowledge of template (and reaction type)

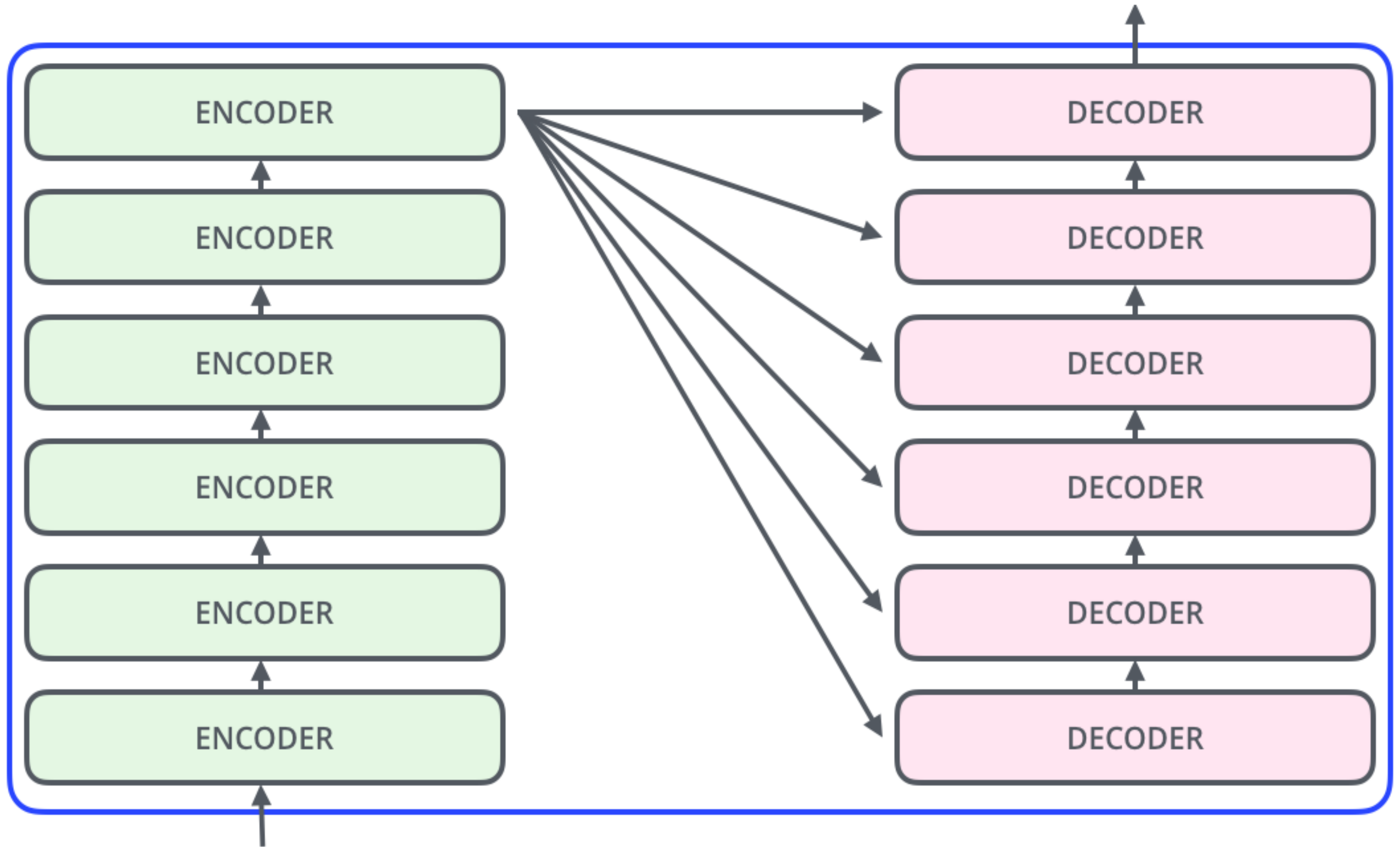
Disadvantages:

- Incomplete coverage of test set
- Cannot generalize outside rule set

- Coverage vs scalability tradeoff
- Relevance: Rules for a given molecule

Prior Work: Template-Free

Cc1cccc(C#C[Si](C)(C)C)n1.Cn1nccc1-c1ccc(Br)cc1



Cc1cccc(C#Cc2ccc(-c3ccnn3C)cc2)n1

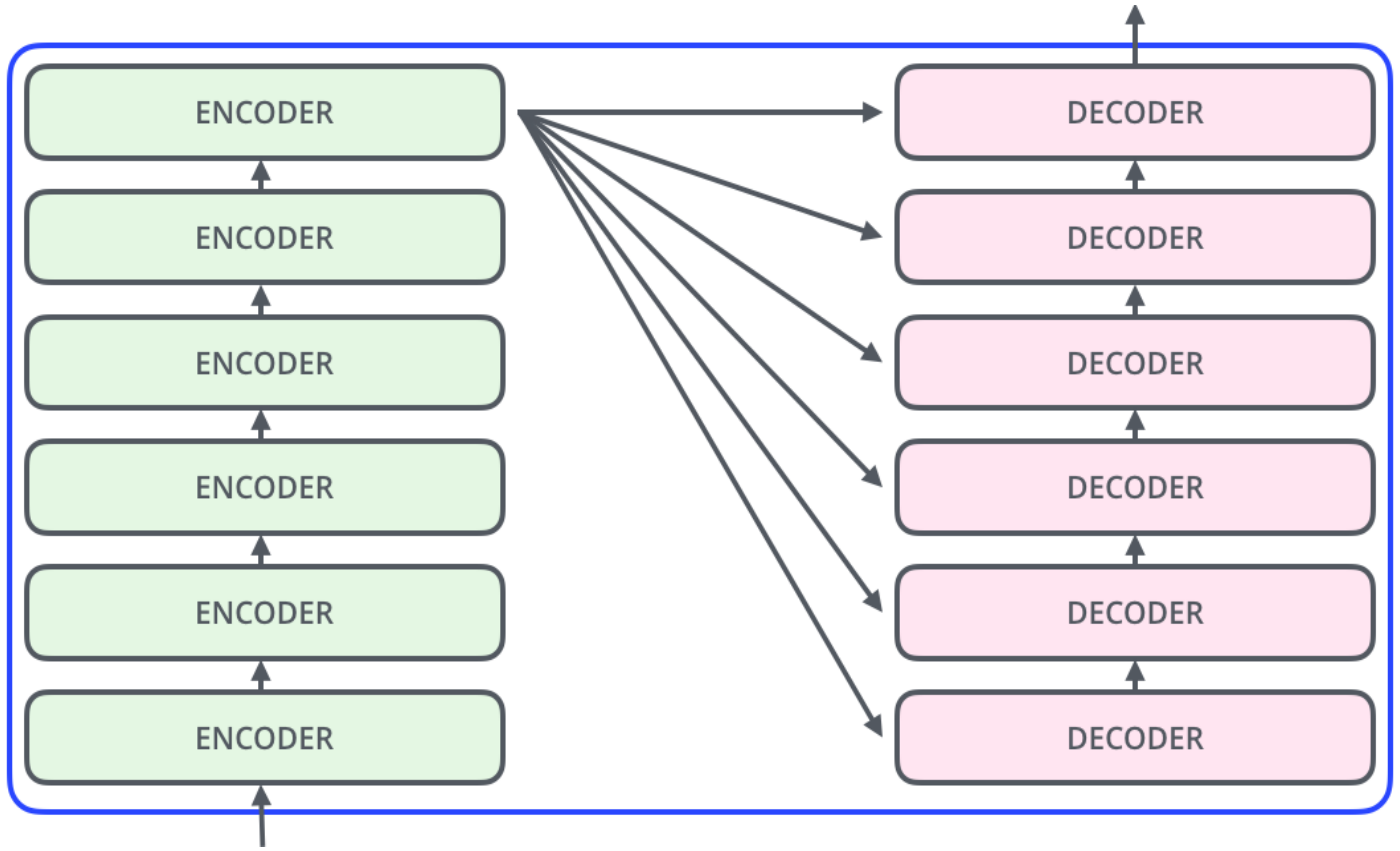
Schwaller et al. (2019)
Zheng et al. (2019)
Chen et al. (2020)

MIT and ETH zürich

Figure: http://jalammar.github.io/illustrated-transformer/

Prior Work: Template-Free

Cc1cccc(C#C[Si](C)(C)C)n1.Cn1nccc1-c1ccc(Br)cc1



Cc1cccc(C#Cc2ccc(-c3ccnn3C)cc2)n1

Schwaller et al. (2019)
Zheng et al. (2019)
Chen et al. (2020)

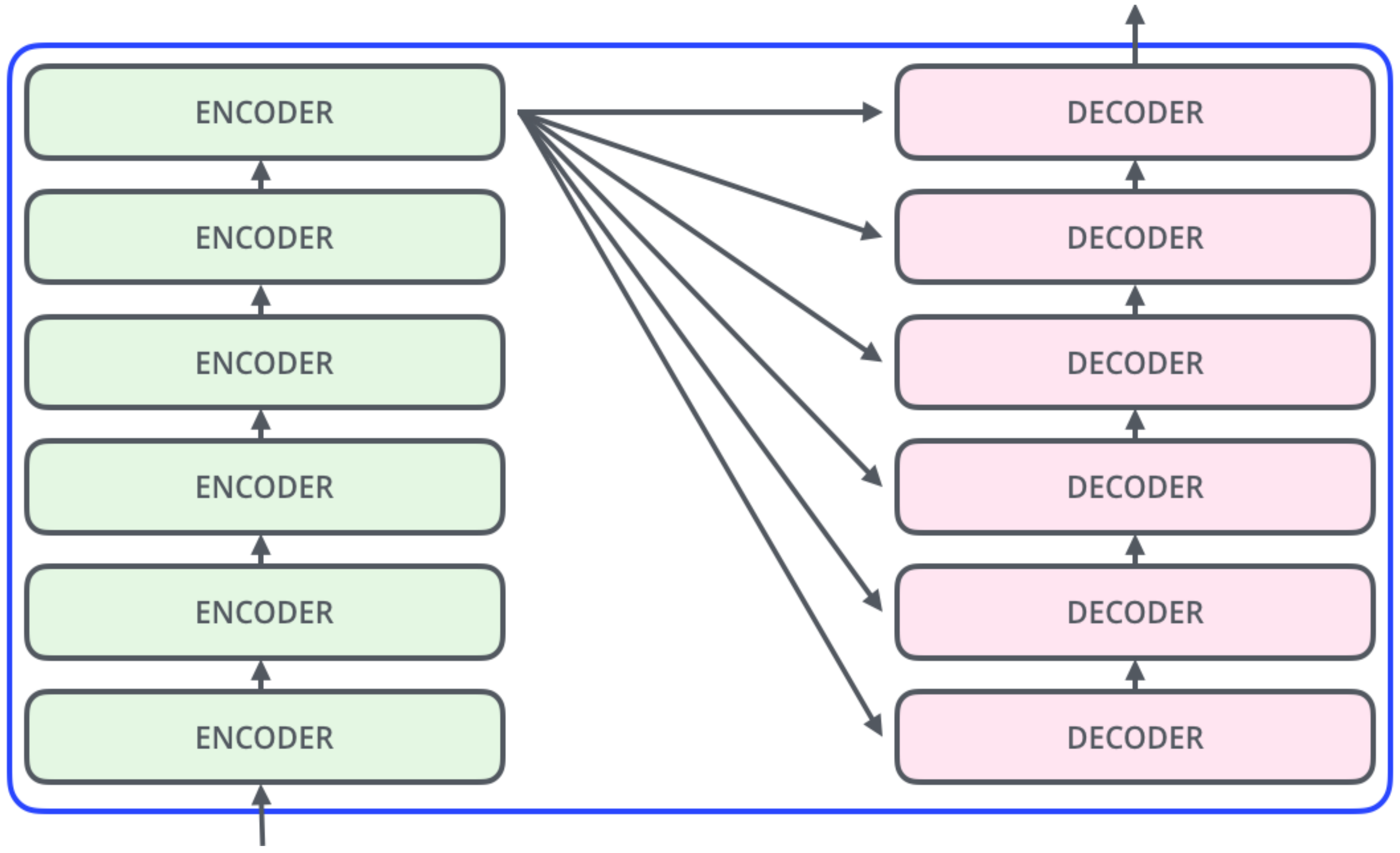
- Discover reaction rules automatically

MIT and ETH zürich

Figure: http://jalammar.github.io/illustrated-transformer/

Prior Work: Template-Free

Cc1cccc(C#C[Si](C)(C)C)n1.Cn1nccc1-c1ccc(Br)cc1



Cc1cccc(C#Cc2ccc(-c3ccnn3C)cc2)n1

Schwaller et al. (2019)
Zheng et al. (2019)
Chen et al. (2020)

Advantages:

- Flexibility in learning transformations

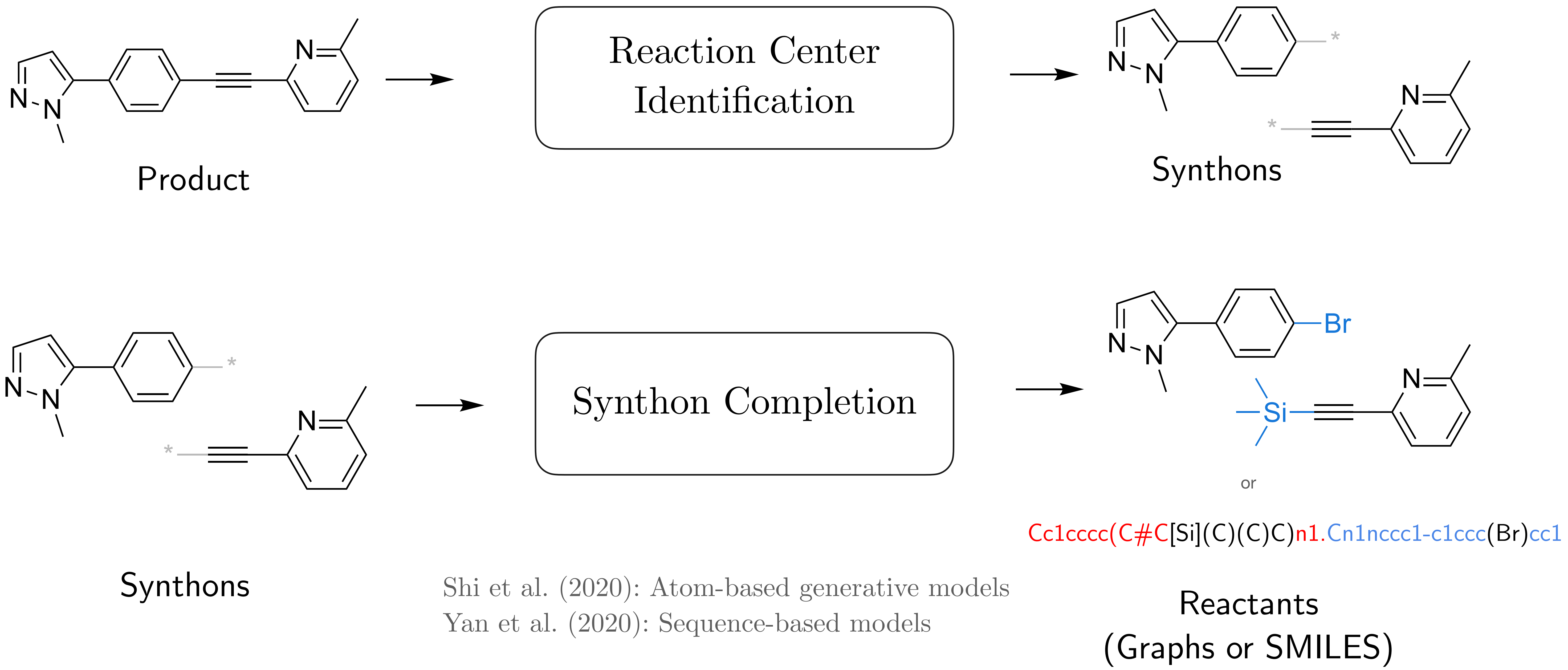
Disadvantages:

- Poor interpretability
- Fails to utilize conserved substructures

- Discover reaction rules automatically

Figure: http://jalammar.github.io/illustrated-transformer/

Prior Work: Semi-Template-Based



Product → Reaction Center Identification → Synthons

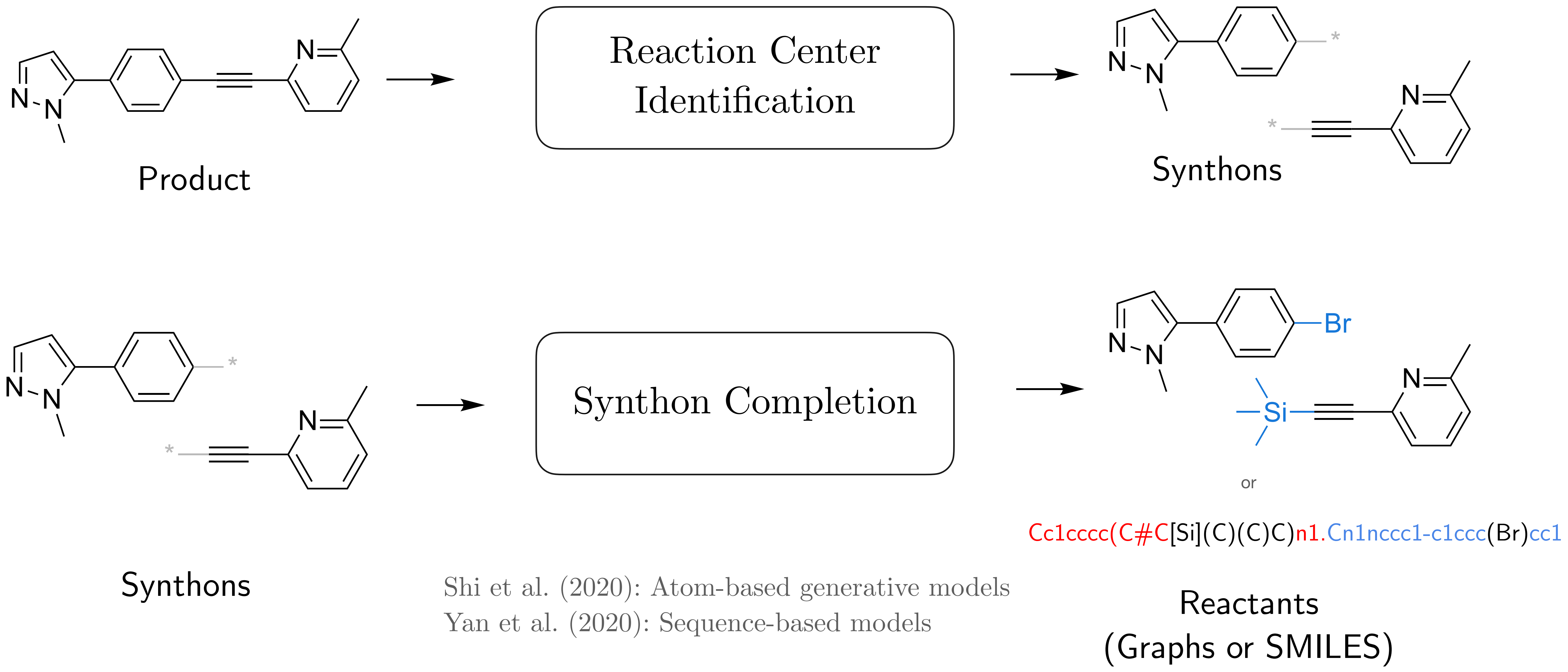
Synthons → Synthon Completion → Reactants (Graphs or SMILES)

or

Cc1cccc(C#C[Si](C)(C)C)n1.Cn1nccc1-c1ccc(Br)cc1

Shi et al. (2020): Atom-based generative models
Yan et al. (2020): Sequence-based models

Prior Work: Semi-Template-Based



Product → Reaction Center Identification → Synthons

Synthons → Synthon Completion → Reactants (Graphs or SMILES)

Shi et al. (2020): Atom-based generative models
Yan et al. (2020): Sequence-based models

Advantages:

1. Closer to a chemist's intuition
2. Improved interpretability

Disadvantages:

1. Fails to utilize conserved substructures in synthon completion

Motivation

Build a retrosynthesis model to identify and utilize conserved substructures

Motivation

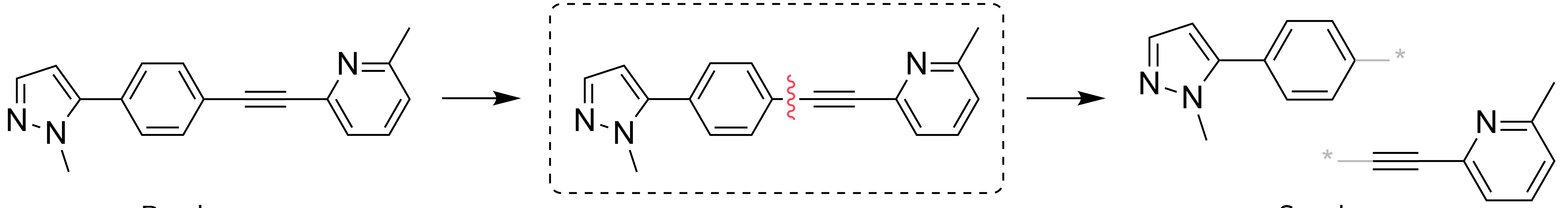
Build a retrosynthesis model to identify and utilize conserved substructures

Advantages

- *Interpretability* - Captures a chemists workflow about retrosynthesis
- *Efficiency* - More efficient use of the data, by not generating/completing molecules from scratch
- *Generalization* - Stronger inductive biases, fewer invalid suggestions

Proposed Formulation

Edit Prediction



Product

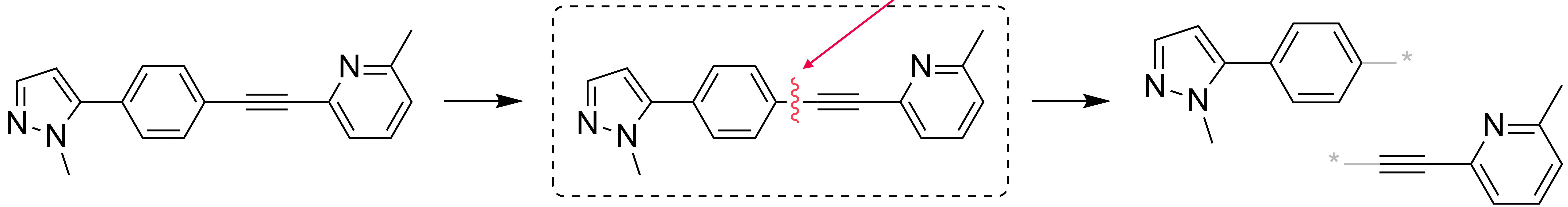
Synthons

Proposed Formulation

Bond Edit

change in bond order

Edit Prediction



Product

Synthons

Synthons

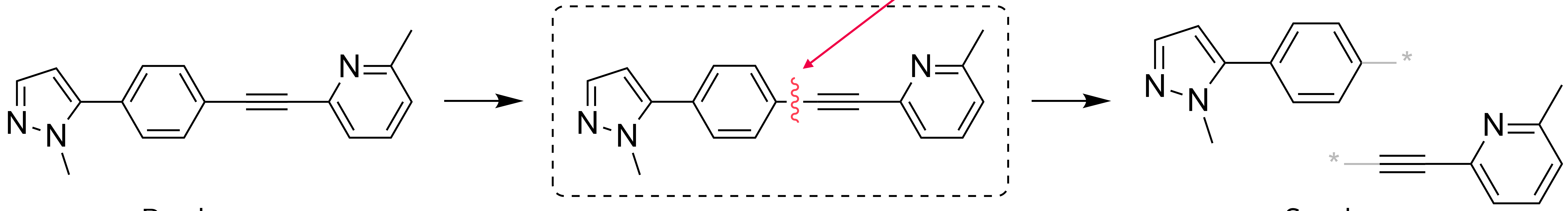
disconnected components after applying the edit

Proposed Formulation

Bond Edit

change in bond order

Edit Prediction



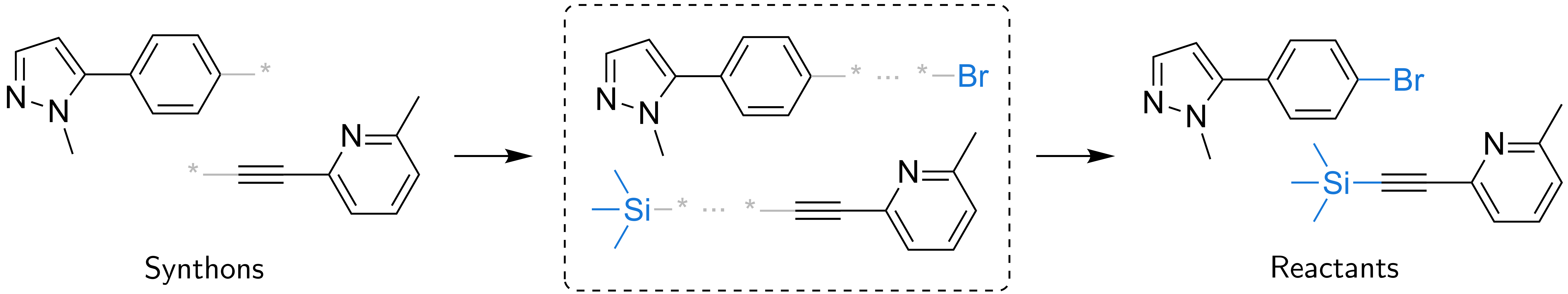
Product

Synthons

Synthons

disconnected components after applying the edit

Synthon Completion



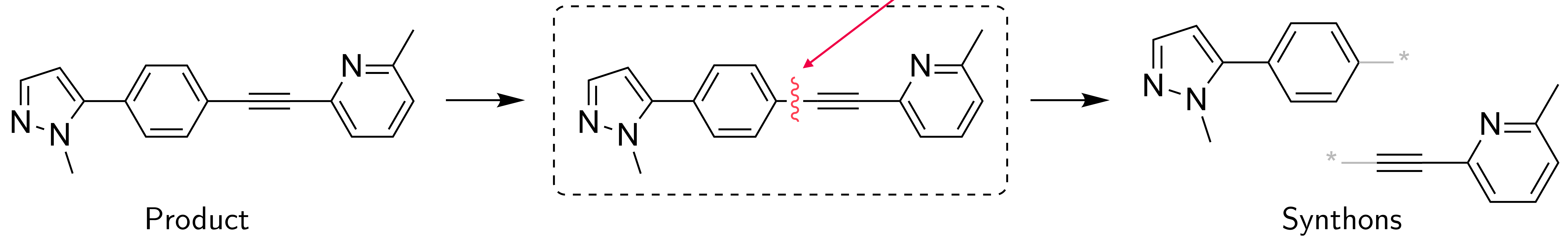
Synthons

Reactants

Proposed Formulation

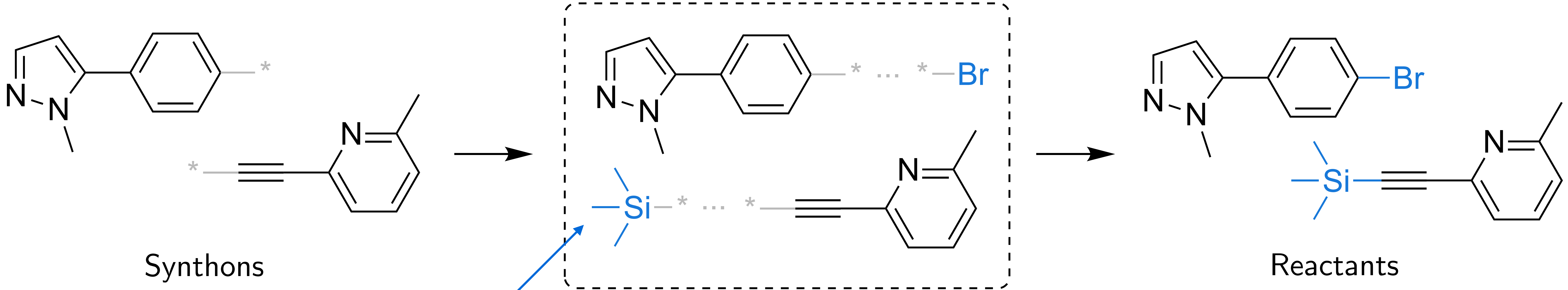
Bond Edit
change in bond order

Edit Prediction



Synthons
disconnected components after applying the edit

Synthon Completion



Leaving Groups
subgraphs added to synthons to produce reactants

and

Edit Prediction

Extracting Edits

Compare atom-maps of products and reactants to identify atoms/bonds undergoing a change

Edit Prediction

Extracting Edits

Compare atom-maps of products and reactants to identify atoms/bonds undergoing a change

Initial Prediction Task

- Use atom and bond representations to predict scores for possible edits
- Allowed edits:
  Whether the hydrogen atom count for a given atom changes (0 or 1)
  Change in the bond type of a given bond (5 possible values)

Edit Prediction

Extracting Edits

Compare atom-maps of products and reactants to identify atoms/bonds undergoing a change

Initial Prediction Task

- Use atom and bond representations to predict scores for possible edits
- Allowed edits:
  Whether the hydrogen atom count for a given atom changes (0 or 1)
  Change in the bond type of a given bond (5 possible values)

Edit Correction

- Leverage dependencies between edits to update initial edit scores
  e.g. aromatic rings are stable and tend to remain unchanged
- LSTM style update on line-graph based representations

Train with cross-entropy loss over possible edits in the molecule

Synthon Completion

Leaving Group Vocabulary Extraction

- Extract subgraphs based on atom-maps only present in reactants
- Small vocabulary size covers 99.7% of the test set

Synthon Completion

Leaving Group Vocabulary Extraction

- Extract subgraphs based on atom-maps only present in reactants
- Small vocabulary size covers 99.7% of the test set

Classification problem instead of a generative one

- Predict the correct leaving group given a synthon
- Teacher forcing during training

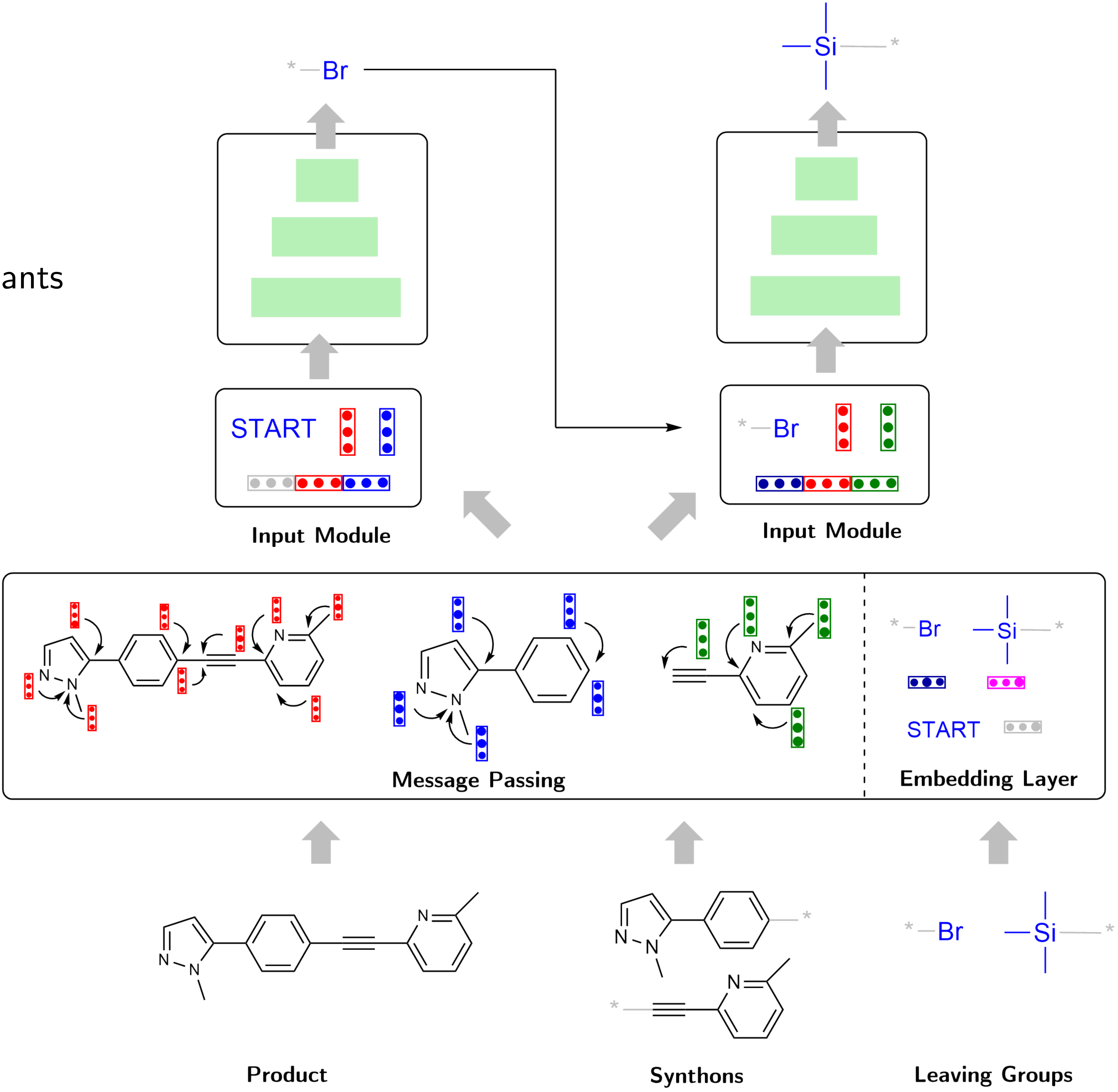
Synthon Completion

Leaving Group Vocabulary Extraction

- Extract subgraphs based on atom-maps only present in reactants
- Small vocabulary size covers 99.7% of the test set

Classification problem instead of a generative one

- Predict the correct leaving group given a synthon
- Teacher forcing during training



MIT and ETH zürich

Experimental Setup

Experimental Setup

Dataset

- USPTO-50k - Standard benchmark dataset
- 50K reactions across 10 reaction classes
- Training/validation/test in a 8:1:1 split (40K train, 5K valid, 5K test)

Experimental Setup

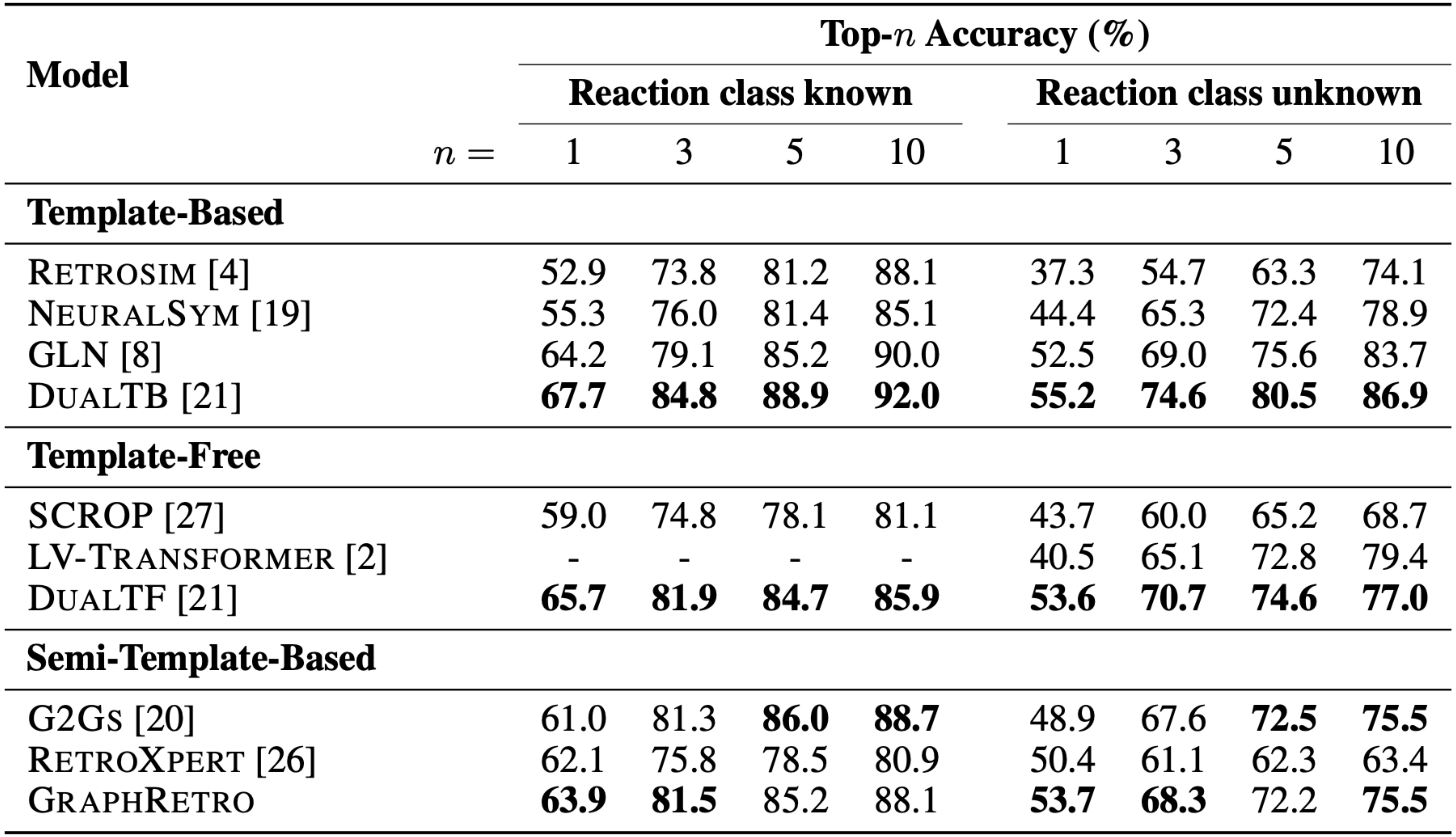
Dataset

- USPTO-50k - Standard benchmark dataset
- 50K reactions across 10 reaction classes
- Training/validation/test in a 8:1:1 split (40K train, 5K valid, 5K test)

Evaluation

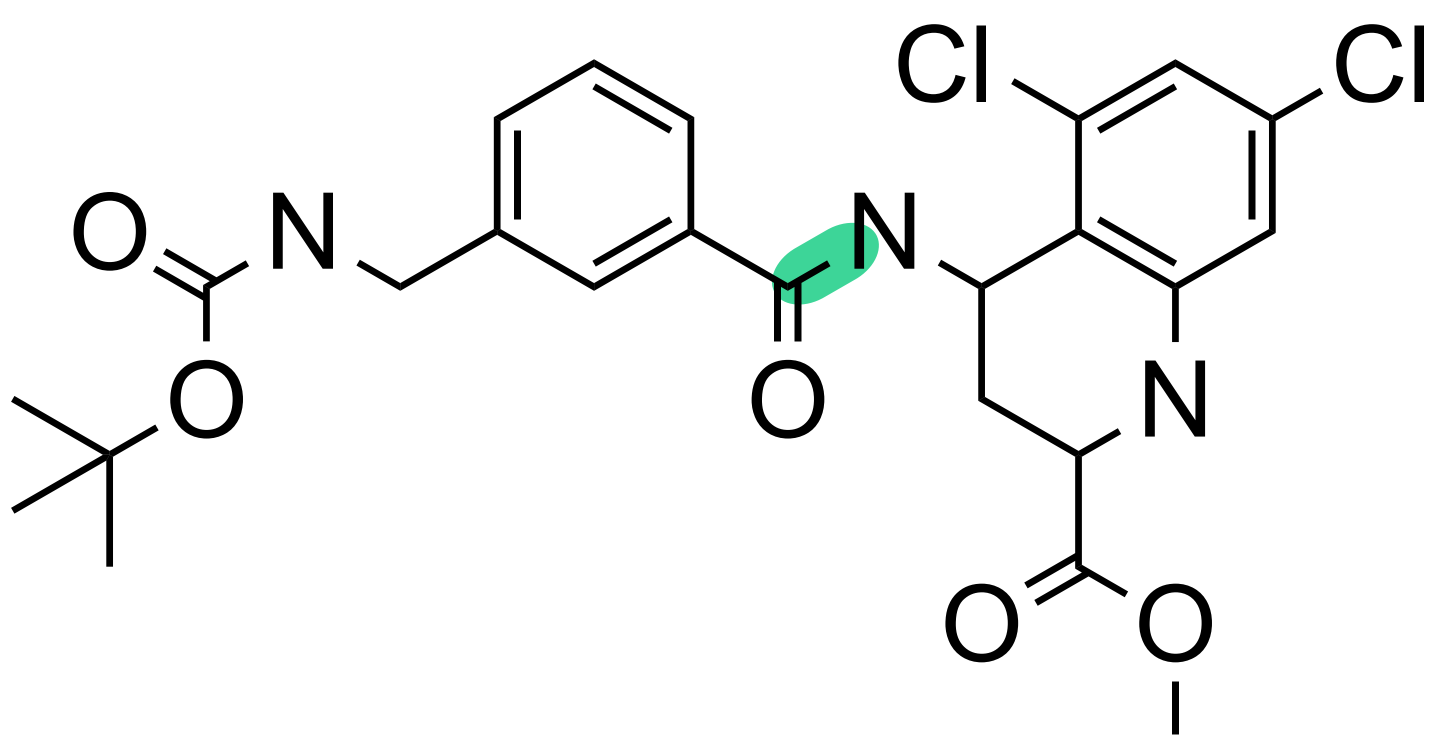
- Top-n accuracy ($n = 1, 3, 5, 10$)
- Compare canonical SMILES of generated reactants to ground truth
- Reaction class known vs unknown

Retrosynthesis Performance

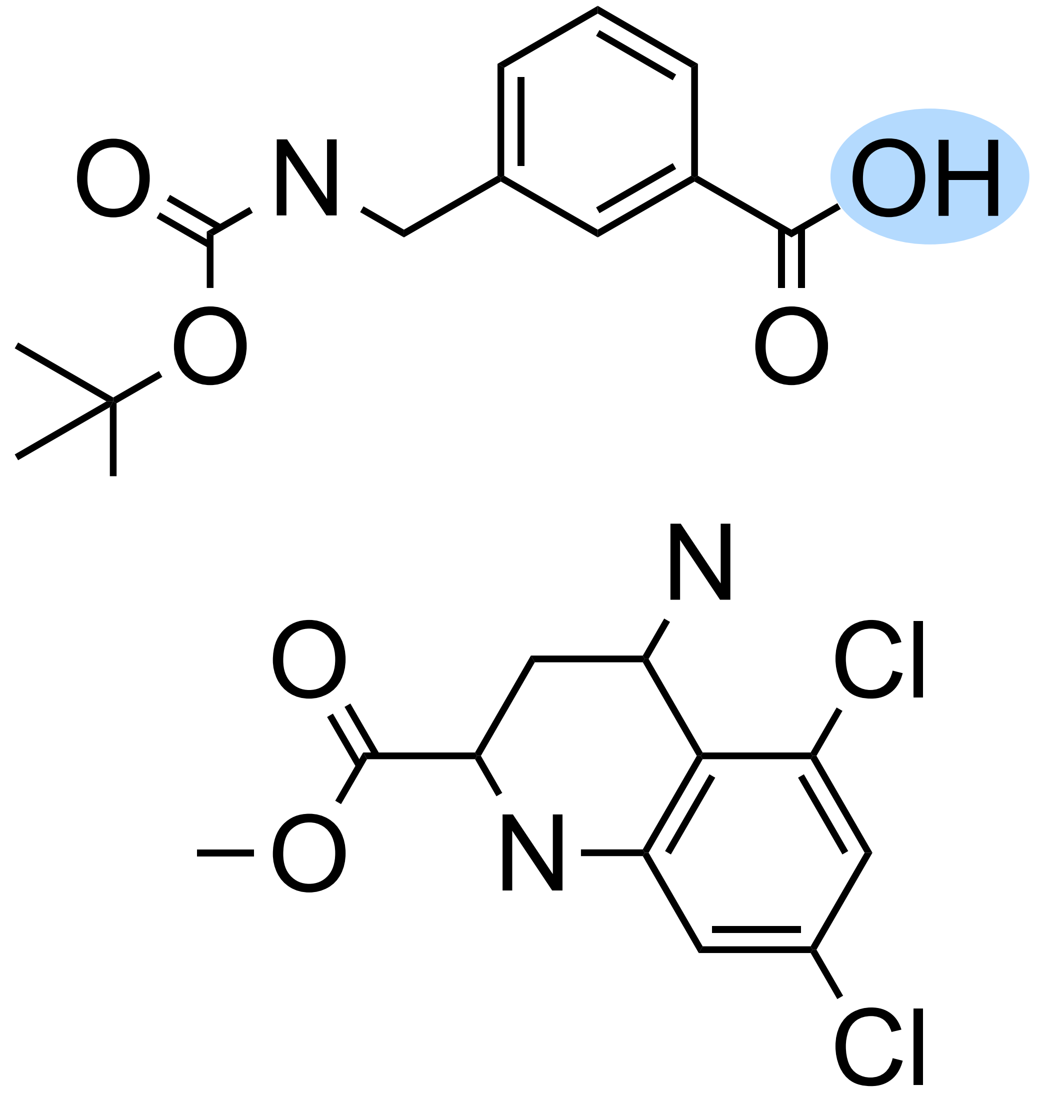


Model		Top-n Accuracy (%)							
		Reaction class known				Reaction class unknown			
	$n =$	1	3	5	10	1	3	5	10
Template-Based									
RetroSim [4]		52.9	73.8	81.2	88.1	37.3	54.7	63.3	74.1
NeuralSym [19]		55.3	76.0	81.4	85.1	44.4	65.3	72.4	78.9
GLN [8]		64.2	79.1	85.2	90.0	52.5	69.0	75.6	83.7
DualTB [21]		**67.7**	**84.8**	**88.9**	**92.0**	**55.2**	**74.6**	**80.5**	**86.9**
Template-Free									
SCROP [27]		59.0	74.8	78.1	81.1	43.7	60.0	65.2	68.7
LV-Transformer [2]		-	-	-	-	40.5	65.1	72.8	79.4
DualTF [21]		**65.7**	**81.9**	**84.7**	**85.9**	**53.6**	**70.7**	**74.6**	**77.0**
Semi-Template-Based									
G2Gs [20]		61.0	81.3	**86.0**	**88.7**	48.9	67.6	**72.5**	**75.5**
RetroXpert [26]		62.1	75.8	78.5	80.9	50.4	61.1	62.3	63.4
GraphRetro		**63.9**	**81.5**	85.2	88.1	**53.7**	**68.3**	72.2	**75.5**

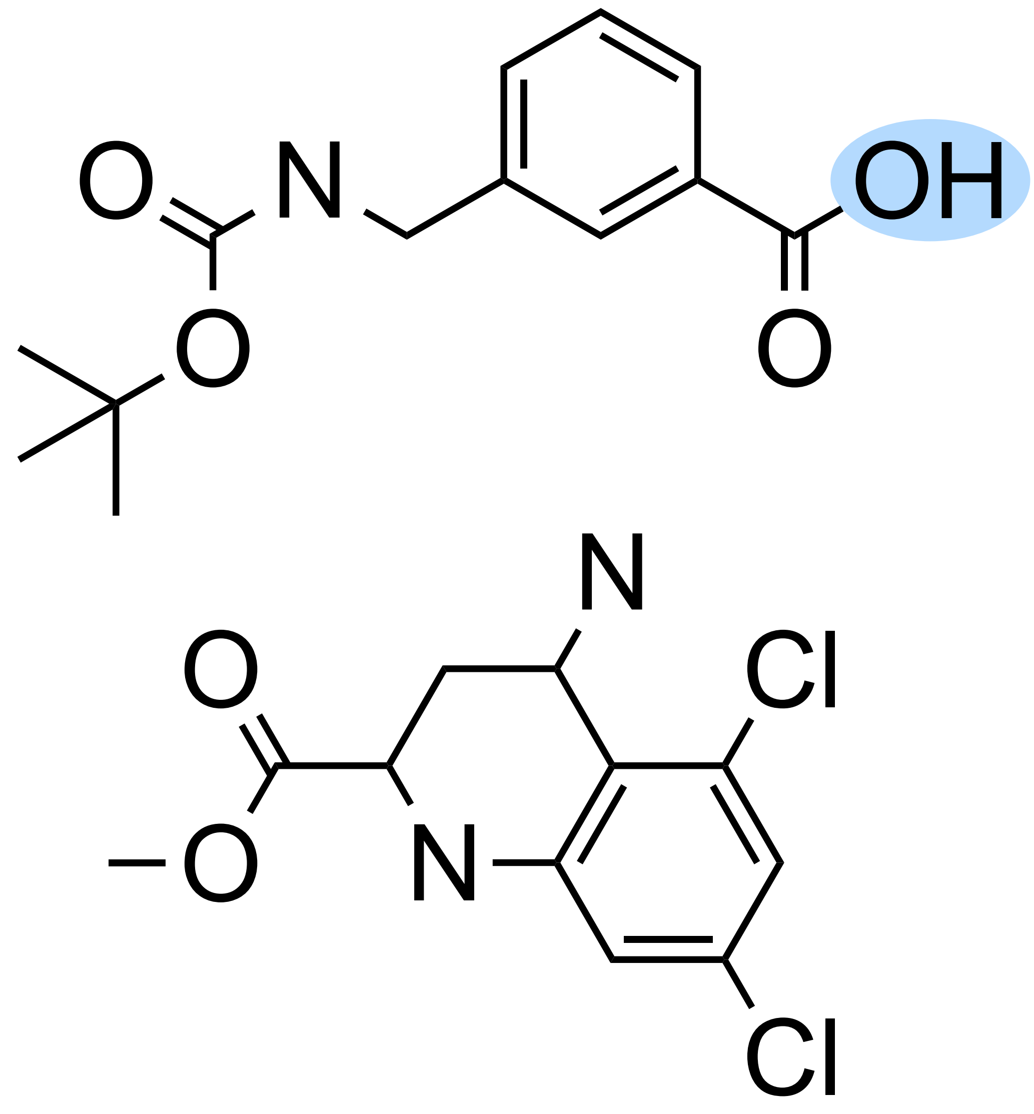
Example Predictions - Correct



Product



True Reactants

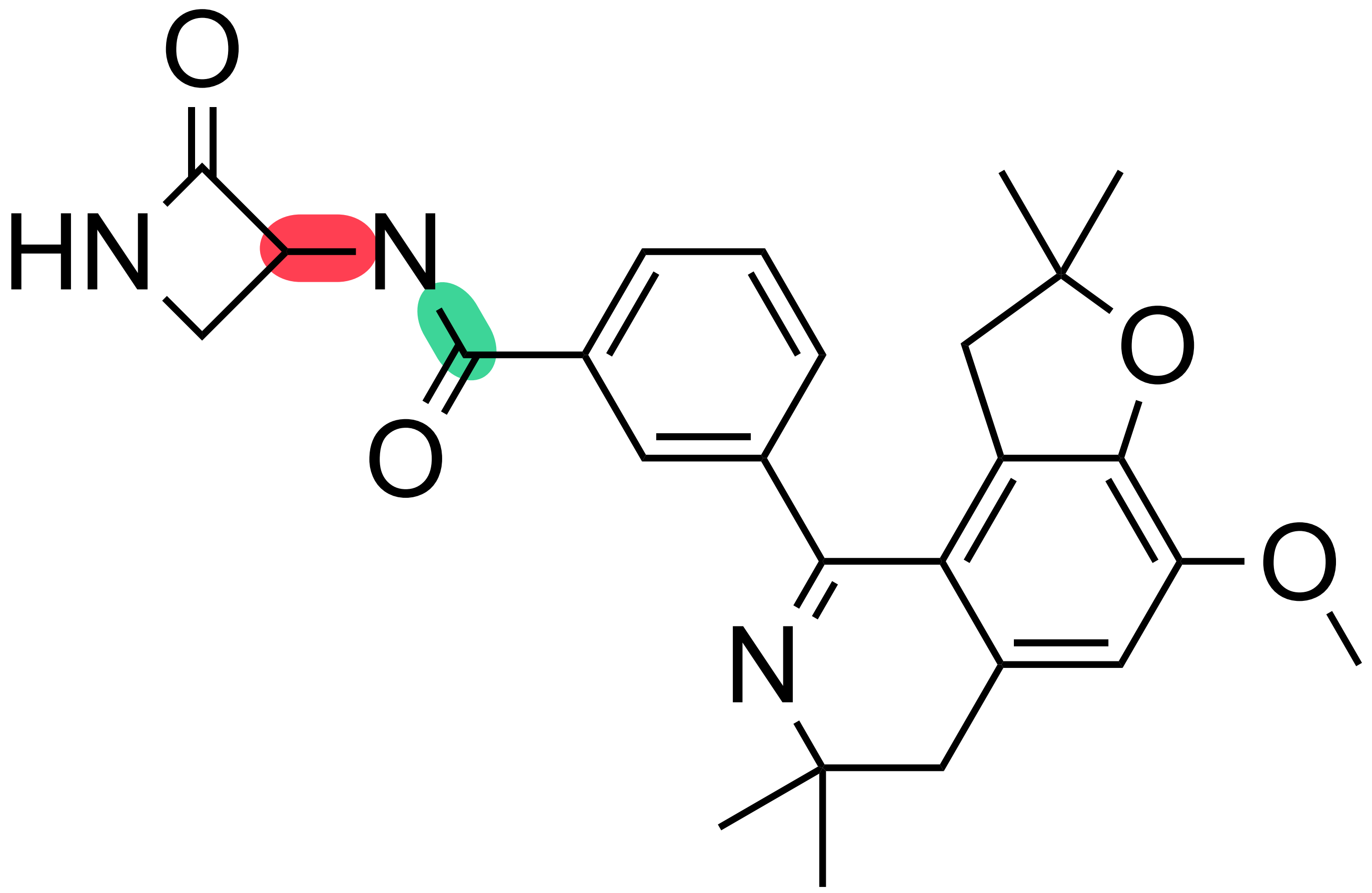


Predicted Reactants

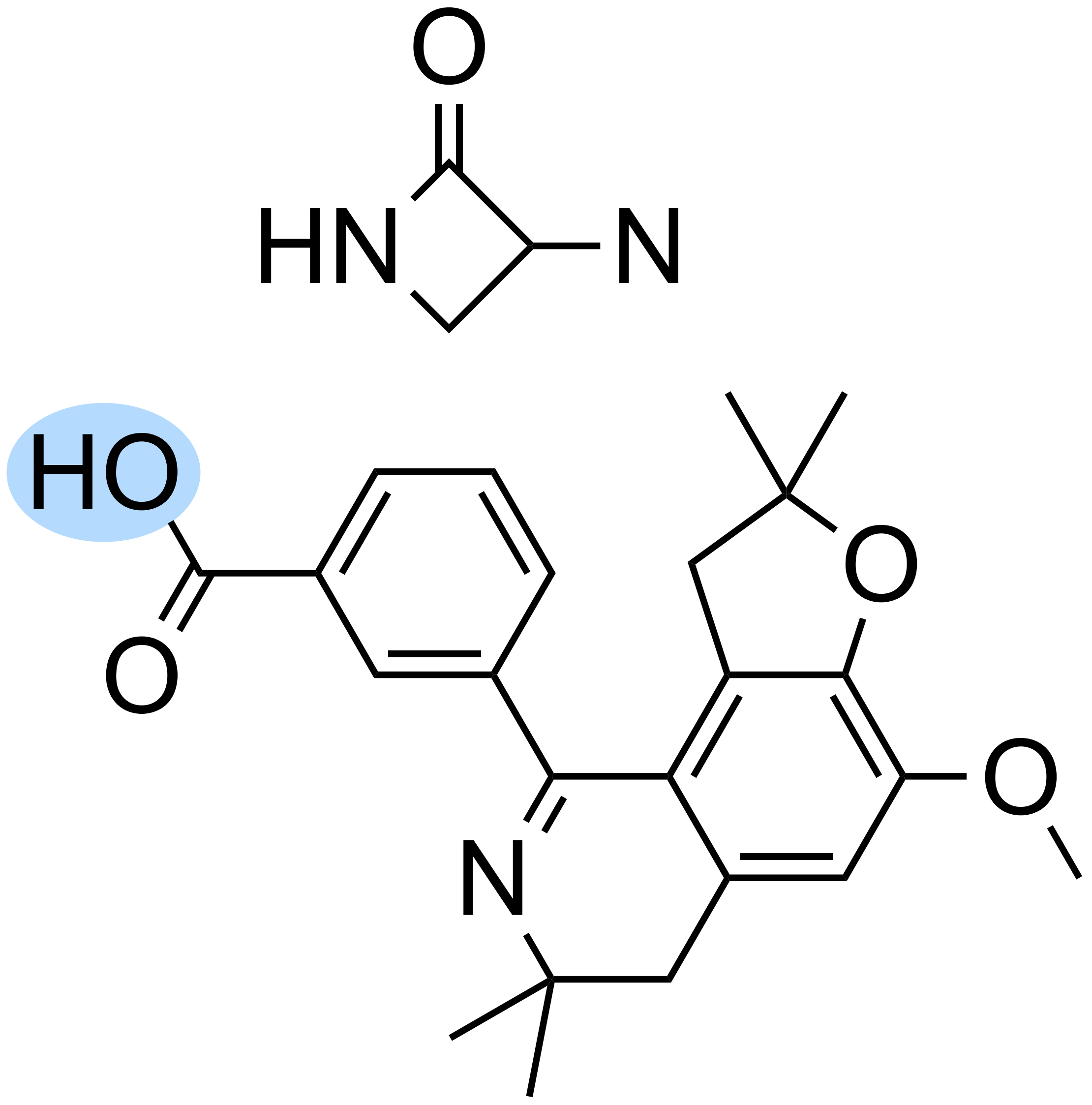
🔵 leaving group 🟢 correct edit

Example Predictions - Incorrect

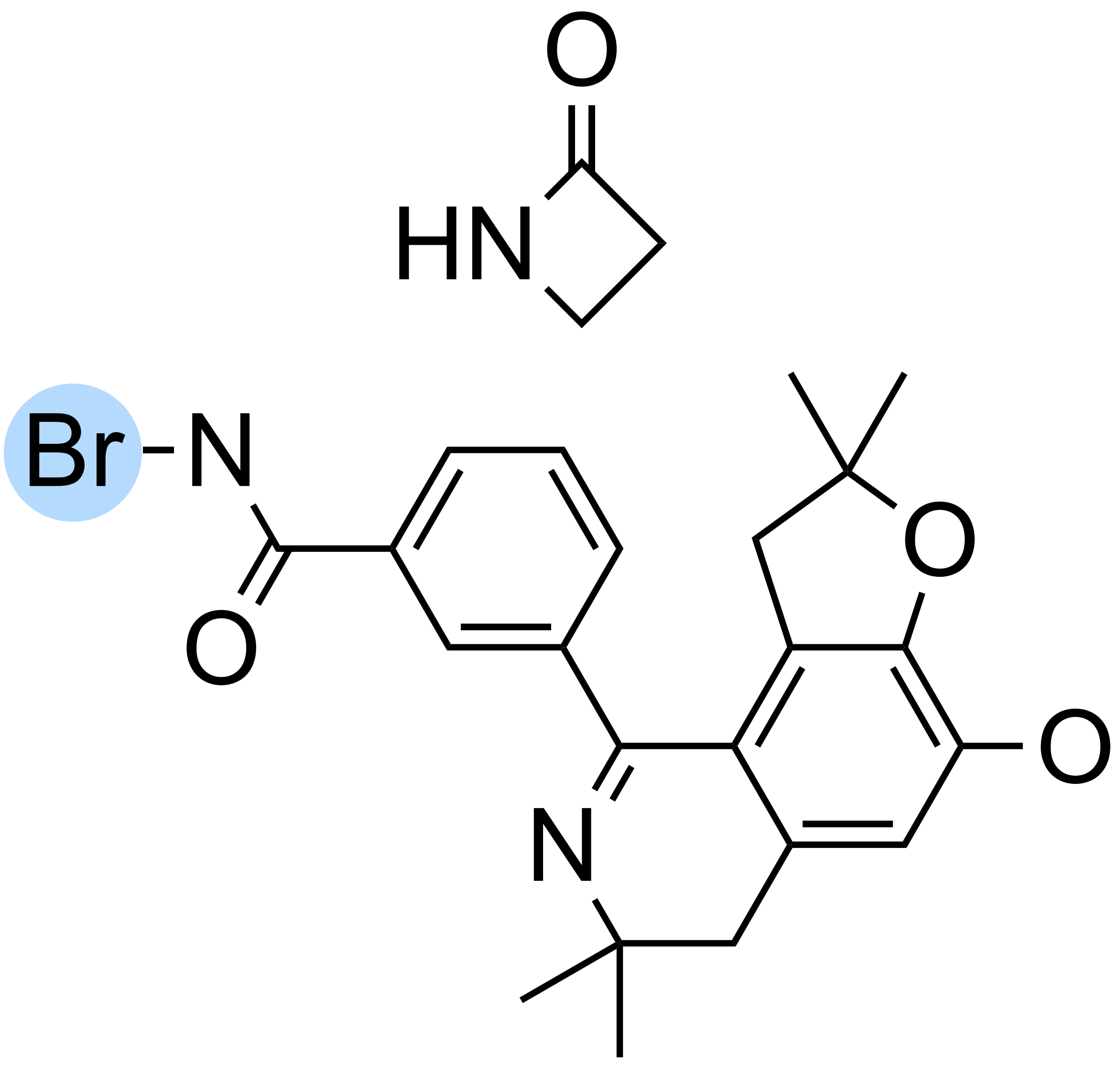
Incorrect edit, leaving groups predicted can't salvage the prediction



Product



True Reactants

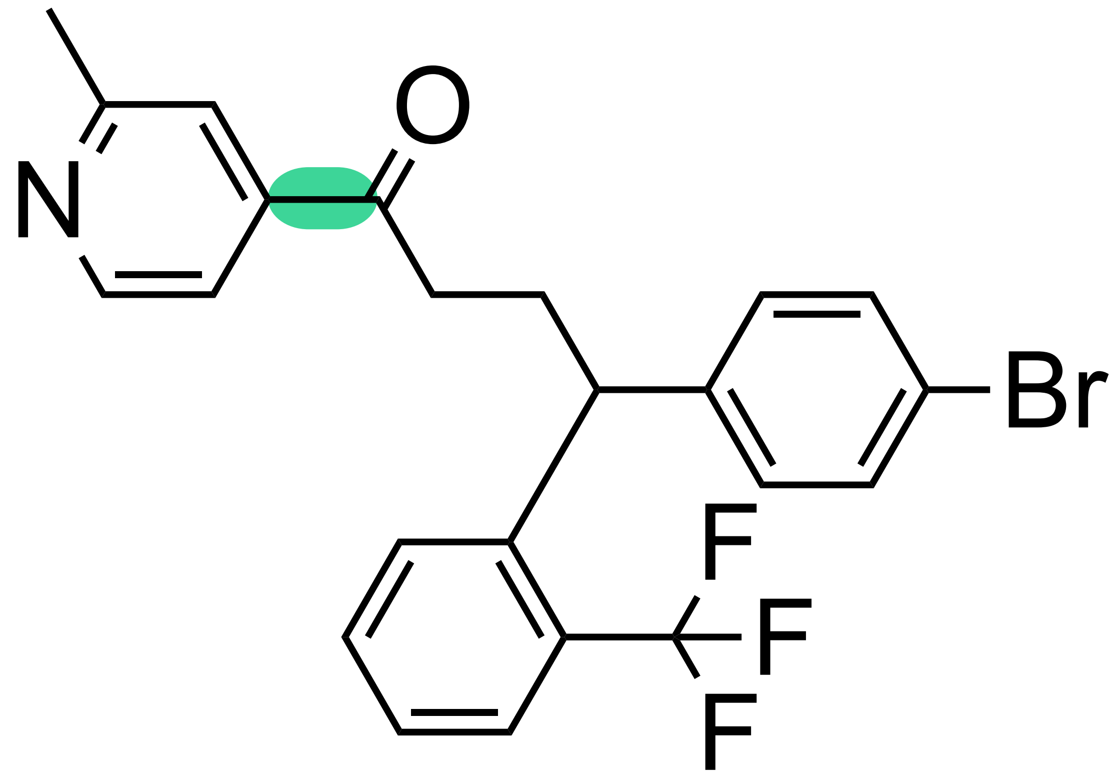


Predicted Reactants

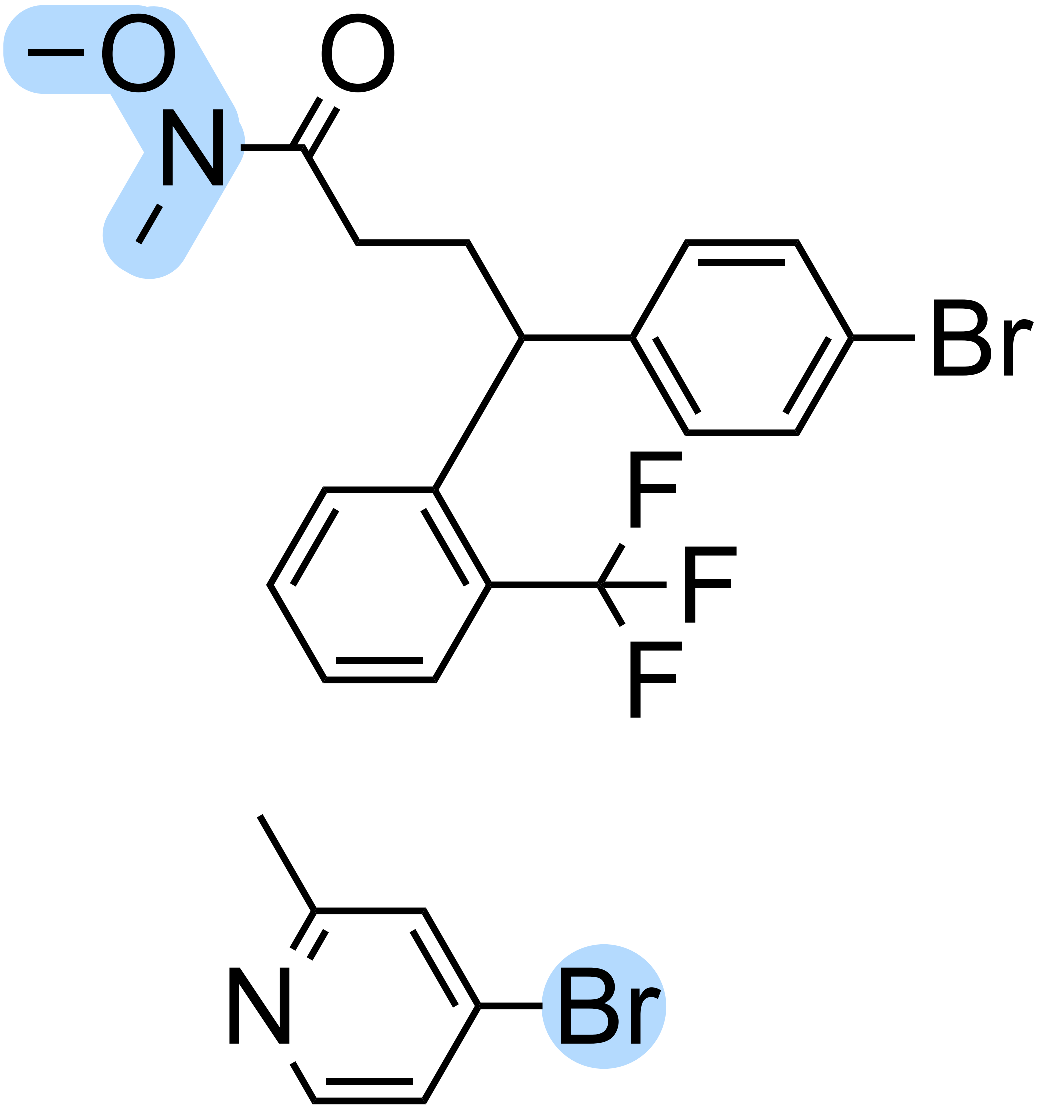
🟦 leaving group 🟩 correct edit 🟥 incorrect edit

Example Predictions - Incorrect

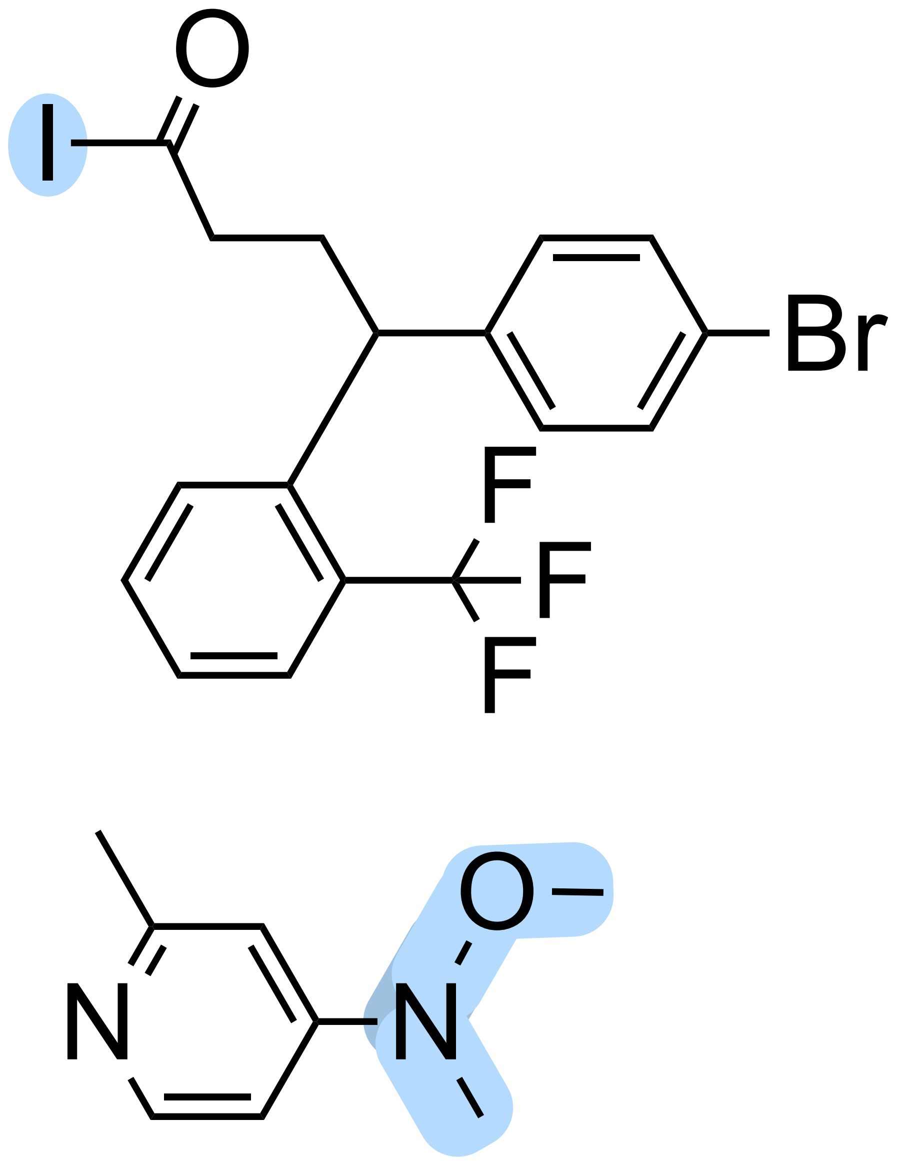
Correct edit, but flipped leaving groups



Product



True Reactants



Predicted Reactants

🟦 leaving group 🟩 correct edit

Summary

- Propose a semi-template based method for retrosynthesis prediction
- Improves top-1 accuracy over previous semi-template methods and template-free methods

Summary

- Propose a semi-template based method for retrosynthesis prediction
- Improves top-1 accuracy over previous semi-template methods and template-free methods

Future Work

- Edit prediction performance is a bottleneck to overall performance

Summary

- Propose a semi-template based method for retrosynthesis prediction
- Improves top-1 accuracy over previous semi-template methods and template-free methods

Future Work

- Edit prediction performance is a bottleneck to overall performance
  - Need more chemically meaningful priors and edit correction mechanisms

Summary

- Propose a semi-template based method for retrosynthesis prediction
- Improves top-1 accuracy over previous semi-template methods and template-free methods

Future Work

- Edit prediction performance is a bottleneck to overall performance
  - Need more chemically meaningful priors and edit correction mechanisms
- Extend synthon completion to predict a single reactant from multiple reactants