Supporting Information: Evaluating the Performance of a Transformer-based Organic Reaction Prediction Model

William R. Borrelli* and Joshua Schrier*

Department of Chemistry, Fordham University, 441 E. Fordham Rd, The Bronx, NY, 10458, United States

E-mail: wborrelli@fordham.edu; jschrier@fordham.edu

Data Set Column Legend

All molecular complexity and synthetic accessibility scores were computed in Mathematica 12.1.

Test Sets

- reaction number: denotes the reaction number as indexed in each data set
- source: the source of the test reaction and the page number
- class: the true reaction class for the test reaction
- true reaction smiles: the literature reported reaction SMILES
- notes: notes about the test reaction or a link to the web page where it can be found (if applicable)

Combined Retrosynthesis Results

- reaction number: denotes the reaction number as indexed in each data set
- true rclass: the true reaction class for the test reaction
- true reaction smiles: the literature reported reaction SMILES
- true prod: the literature reported product
- pred reaction smiles: the predicted reaction SMILES from the model output
- pred prod: the predicted product from the model output
- conf: the retrosynthesis confidence from the model output
- projID: the IBM RXN project ID for the test reaction

- retroID: the IBM RXN retrosynthesis ID for the test reaction
- true_deltaCM: the difference in molecular complexity for the literature reported test reaction
- true_deltaSA: the difference in synthetic accessibility for the literature reported test reaction
- retpred_deltaCM: the difference in molecular complexity for the predicted reaction returned from the model
- retpred_deltaSA: the difference in synthetic accessibility for the predicted reaction returned from the model

Combined Forward Synthesis Results

- reaction number: denotes the reaction number as indexed in each data set
- true rclass: the true reaction class for the test reaction
- true reaction smiles: the literature reported reaction SMILES
- true prod: the literature reported product
- pred reaction smiles: the predicted reaction SMILES from the model output
- pred prod: the predicted product from the model output
- conf: the forward synthesis confidence from the model output
- projID: the IBM RXN project ID for the test reaction
- predID: the IBM RXN forward prediction ID for the test reaction
- T1 Correct: boolean if the top-1 prediction was correct

- T2 Correct: boolean if any top-2 predictions were correct
- T3 Correct: boolean if any top-3 predictions were correct
- T4 Correct: boolean if any top-4 predictions were correct
- T5 Correct: boolean if any top-5 predictions were correct
- true_deltaCM: the difference in molecular complexity for the literature reported test reaction
- true_deltaSA: the difference in synthetic accessibility for the literature reported test reaction
- fwdpredo_deltaCM: the difference in molecular complexity for the predicted reaction returned from the model
- retpred_deltaSA: the difference in synthetic accessibility for the predicted reaction returned from the model

Incorrect and High Confidence Forward Predictions

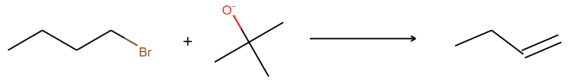
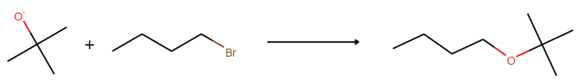
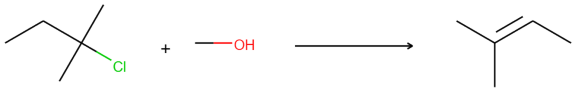
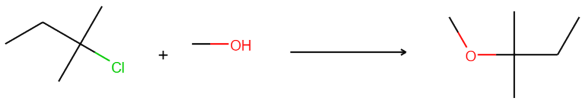
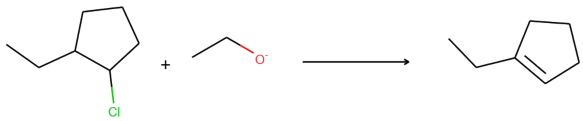
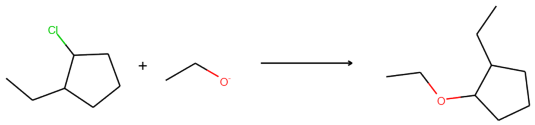
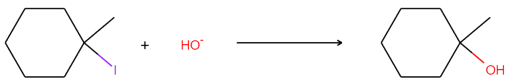
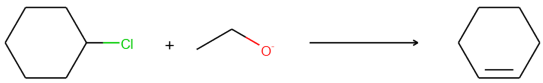
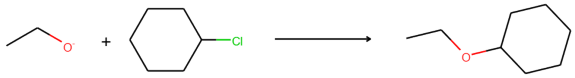
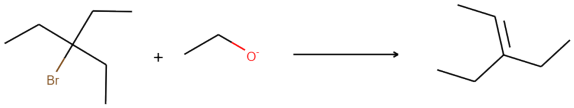
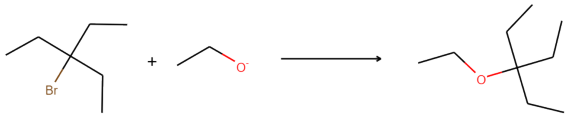
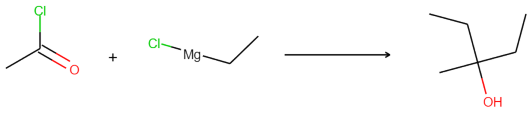
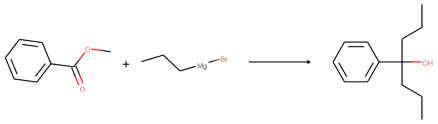
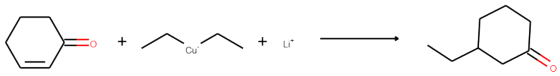
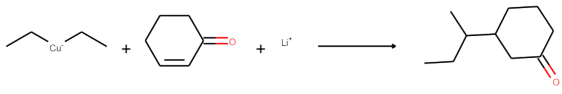
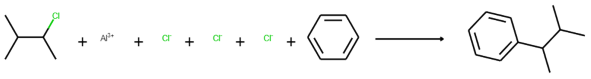
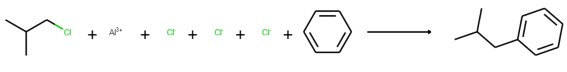
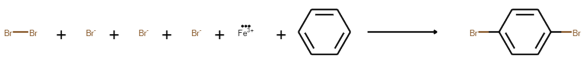
Reaction Number	Literature Reaction	Predicted Reaction
11		
12		
15		
16		
18		
19		
25		Failure
27		Failure
40		
43		
50		
71		

Figure 1: Reactions for All Incorrect and High Confidence Forward Predictions

RXNMapper Map Confidence Distributions

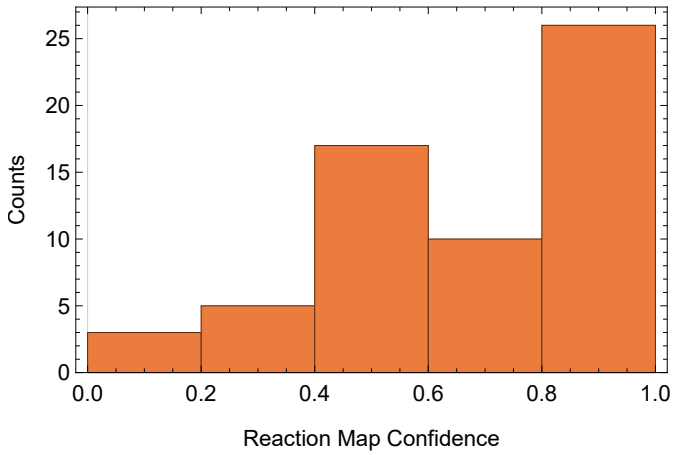


Figure 2: RXNMapper Map Confidences for Top-1 Correct Predictions

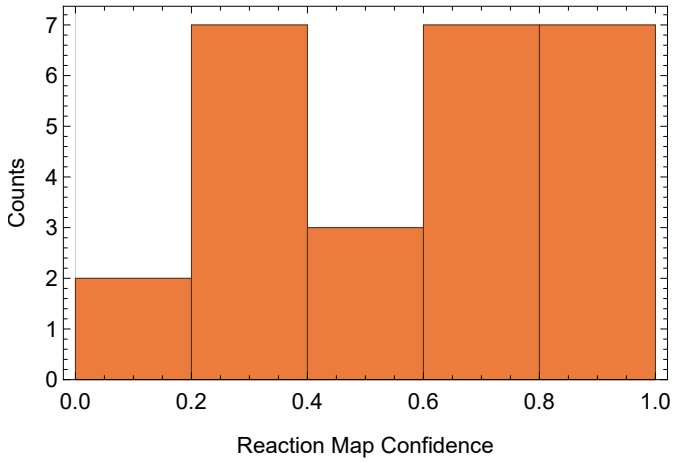


Figure 3: RXNMapper Map Confidences for Top-1 Incorrect Predictions