DiffER: Categorical Diffusion Models for Chemical Retrosynthesis

Sean Current, Ziqi Chen, Daniel Daniel Adu-Ampratwum, Xia Ning, Srinivasan Parthasarathy • current.33@osu.edu • github.com/sfcurre/DiffER

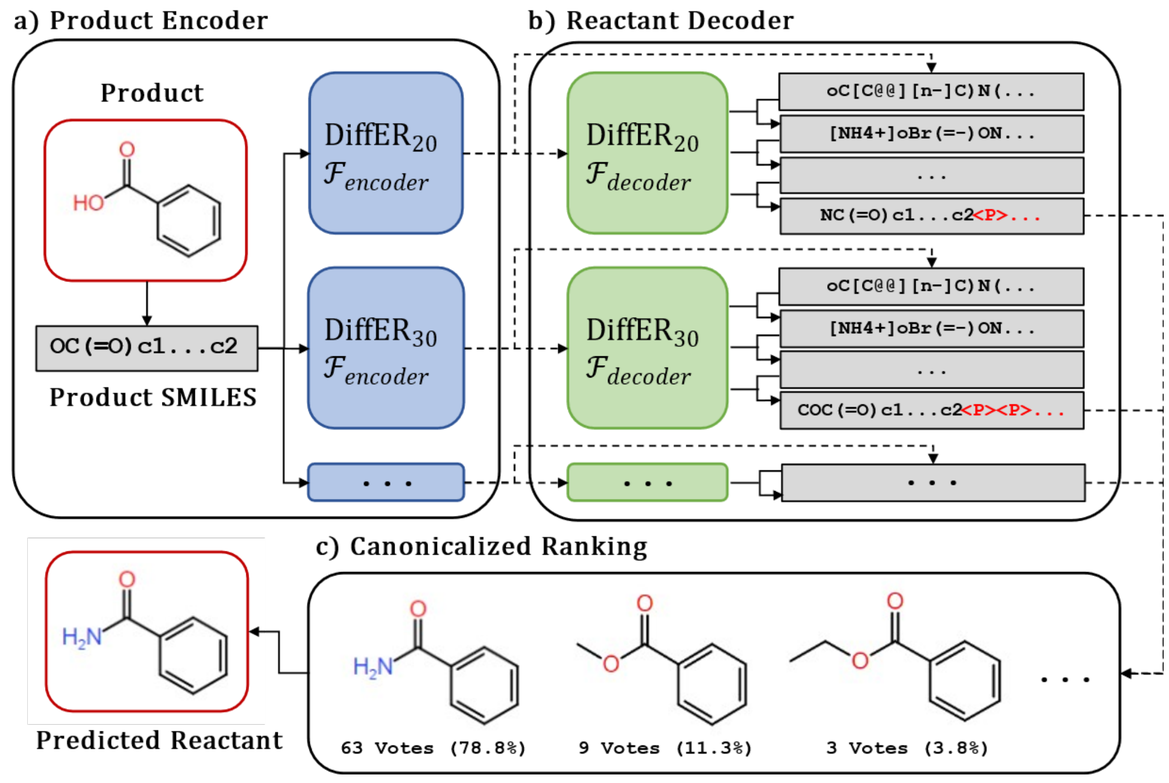
Introduction

Retrosynthesis prediction is a vital step in organic synthesis tasks, particularly for drug discovery and engineering. In forward synthesis prediction, the products of a chemical reaction are predicted from known reactants, while retrosynthesis models the reverse process, predicting possible reactants from a target product. Recent advances in machine learning models for chemical retrosynthesis have taken advantage of transformer architectures originally crafted for natural language tasks. Instead of operating on natural language, these models are set to operate on SMILES encodings of chemical products and reactants. In this work, we propose DiffER, which uses an ensemble of categorical diffusion models to generate reactant SMILES rather than traditional autoregressive decoders. DiffER offers a few advantages over existing approaches:

- DiffER samples from an estimated posterior distribution of the entire reactant SMILES, rather than decoding the reactant SMILES autoregressively token-by-token.
- DiffER's ensemble approach generates a distribution of reactant SMILES, offering measures of confidence and uncertainty in predictions.

These advantages lead to DiffER achieving state-of-the-art top-1 accuracy compared to the baseline methods, and second-best top-3 through top-10 performance. Additionally, our results show that even greater performance can be achieved if the length of the reactant SMILES can be accurately estimated, significantly out-performing all other methods.

Methods



Key Ideas and Implementation Details

Categorical Diffusion

- Iteratively denoises a categorical distribution conditioned on the target product (Hoogeboom, 2021)

$$p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}_0) = \mathcal{C}(\boldsymbol{y}_{t-1}|\boldsymbol{\theta}_{\text{post}}(\boldsymbol{y}_t, \mathcal{F}_{\text{decoder}}(\boldsymbol{y}_t||\mathcal{F}_{\text{encoder}}(\boldsymbol{x}_0))))$$

- Uses both MSE and variational lower bound (VLB) losses

Length Prediction

- Incorporate "change in length" prediction as part of the encoder model (adapted from Ghazvininejad, 2019)
- Append random amount of "padding" tokens to allow model to vary output size to account for variability in SMILES strings

Model Ensembling

- Adjusting the sampling distribution of "padding" tokens produces different models with different capabilities
- We train multiple models and utilize ensemble voting to achieve the best performance

Results

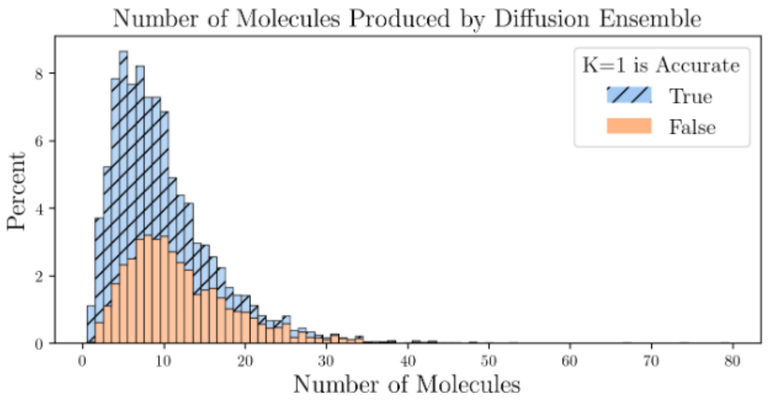
Individual Models

- Theoretical performance is exemplary (oracle)
- Length variant prediction improves performance
- Individual models struggle with diversity
- Fewer different reactants are generated when the most predicted reactant matches the ground truth

Ensemble Model

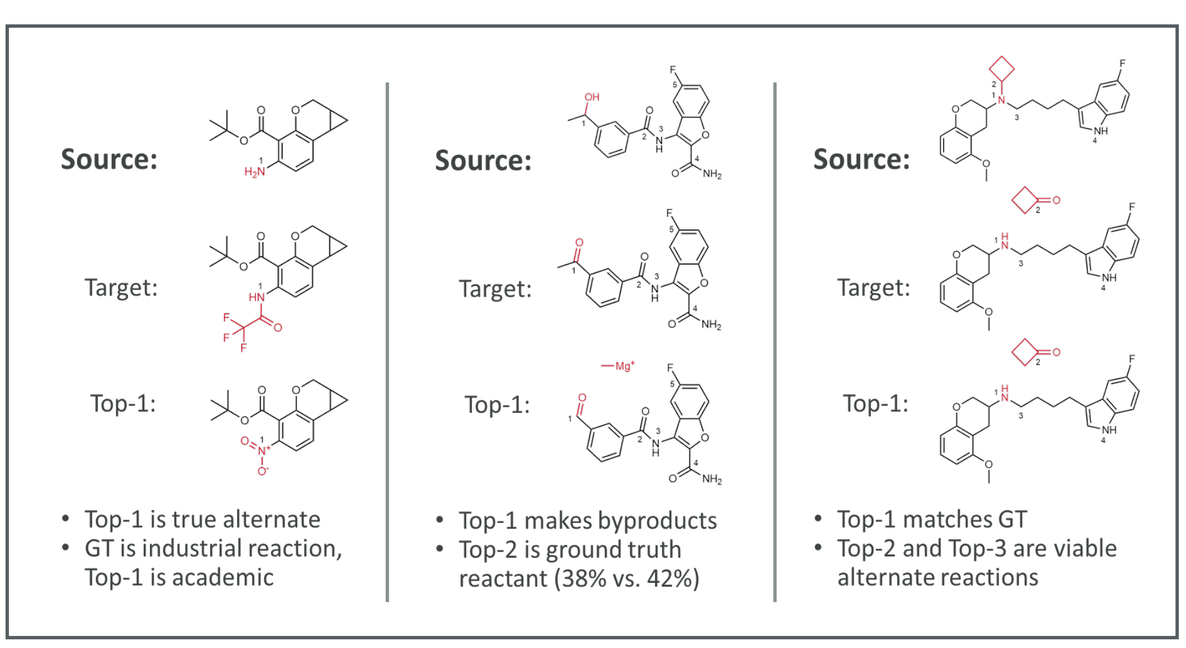
- Best top-1 performance
- Second-best top-3 through top-10 performance
- Performance falls due to lack of output diversity
- Even when predicted reactants do no match patented reactants, DiffER produces chemically plausible alternatives (Case Study)

Model	K=1	3	5	10	Sample Validity	Avg. Num. Reactants
DiffER Ensemble	57.6	79.0	84.1	87.4	100.0	10.0
DiffER $_{20}$	53.2	70.3	72.9	73.6	100.0	3.2
DiffER $_{30}$	55.2	71.5	74.4	75.2	99.9	3.2
DiffER $_{40}$	54.6	72.1	74.4	75.3	99.9	3.2
DiffER $_{50}$	54.9	71.3	73.7	74.4	100.0	3.2
DiffER $_{60}$	55.4	71.7	74.6	75.4	99.9	3.3
DiffER $_{70}$	54.3	71.1	73.5	74.4	99.6	3.2
DiffER $_{80}$	54.6	71.9	74.4	75.1	99.6	3.3
DiffER $_{90}$	54.5	71.6	74.2	74.9	99.8	3.3
Baseline Length	40.4	55.9	58.8	59.9	99.9	3.8
Oracle Length	77.0	88.1	89.5	90.0	99.7	2.8



Category	Model	K=1	3	5	10
Template-based	Retrosim	37.3	54.7	63.3	74.1
	Neuralsym	44.4	65.3	72.4	78.9
	GLN	52.5	69.0	75.6	83.7
	LocalRetro	**53.4**	**77.5**	**85.9**	**92.4**
Semi-template	G2Gs	48.9	67.6	72.5	75.5
	GraphRetro	53.7	68.3	72.2	75.5
	RetroXpert	50.4	61.1	62.3	63.4
	RetroPrime	51.4	70.8	74.0	76.1
	G^{2}Retro	**53.9**	**74.6**	**80.7**	**86.6**
Template-free	Seq2Seq	37.4	52.4	57.0	61.7
	Levenshtein	41.5	48.1	50.0	51.4
	GTA	51.1	67.6	74.8	81.6
	Graph2SMILES	51.2	66.3	70.4	73.9
	Dual-TF	53.3	69.7	73.0	75.0
	MEGAN	48.1	70.7	78.4	86.1
	Chemformer	54.3	-	62.3	63.0
	Retroformer	53.2	71.1	76.6	82.1
	Tied transformer	47.1	67.2	73.5	78.5
	R-SMILES	56.3	**79.2**	**86.2**	**91.0**
	DiffER	**57.6**	79.0	84.1	87.4

Case Study



- Top-1 is true alternate
- GT is industrial reaction, Top-1 is academic

- Top-1 makes byproducts
- Top-2 is ground truth reactant (38% vs. 42%)

- Top-1 matches GT
- Top-2 and Top-3 are viable alternate reactions

Conclusion

By utilizing an ensemble of categorical diffusion models over traditional autoregressive methods, DiffER achieves state-of-the-art top-1 accuracy for single step retrosynthesis prediction, and second-best top-3 through top-10 performance. Unfortunately, DiffER suffers from a lack of sample diversity, as the same molecule may be sampled multiple times from the estimated posterior. In future work, we aim to improve the length-prediction component of the model, as well as implement reinforcement learning fine-tuning approaches to further improve diffusion performance.

References

1. Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. Advances in Neural Information Processing Systems, 34:12454–12465, 2021.
2. Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. arXiv preprint arXiv:1904.09324, 2019.

This project was made possible, in part, by support from the National Science Foundation grant no. IIS-2133650 and the National Library of Medicine grant no. 1R01LM014385.
Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agency.