Supporting information

Assessing the Extrapolation Capability of Template-free Retrosynthesis Models

Jonghwi Choe[1]‡, Shuan Chen[1,2]‡, and Yousung Jung[1,2,3*]

[1] *Department of Chemical and Biological Engineering, Seoul National University, 1,*

Gwanak-ro, Gwanak-gu, Seoul, 08826, South Korea

[2] *Institute of Chemical Processes, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826, South Korea.*

[3] *Interdisciplinary Program in Artificial Intelligence, Seoul National University, 1,*

Gwanak-ro, Gwanak-gu, Seoul, 08826, South Korea.

† Electronic Supplementary Information (ESI) available: [details of any supplemen-tary information available should be included here]. See DOI: 00.0000/00000000.

‡ These authors contributed equally to this work' as above using the symbols: ‡

*Email: yousung.jung@snu.ac.kr

S1. Dataset Composition and Template Distribution across Data Splits

Table S1. Dataset statistics and template counts for training, validation, and test sets in both the USPTO-50k and USPTO-480k datasets.

	Dataset	Train	Validation	Test
# Reactions	USPTO-50k	39,982	5,016	5,018
	USPTO-480k	383,784	46,519	48,730
# Templates	USPTO-50k	45	40	835
	USPTO-480k	232	667	19,321

S2. Template Popularity Distribution in USPTO Datasets

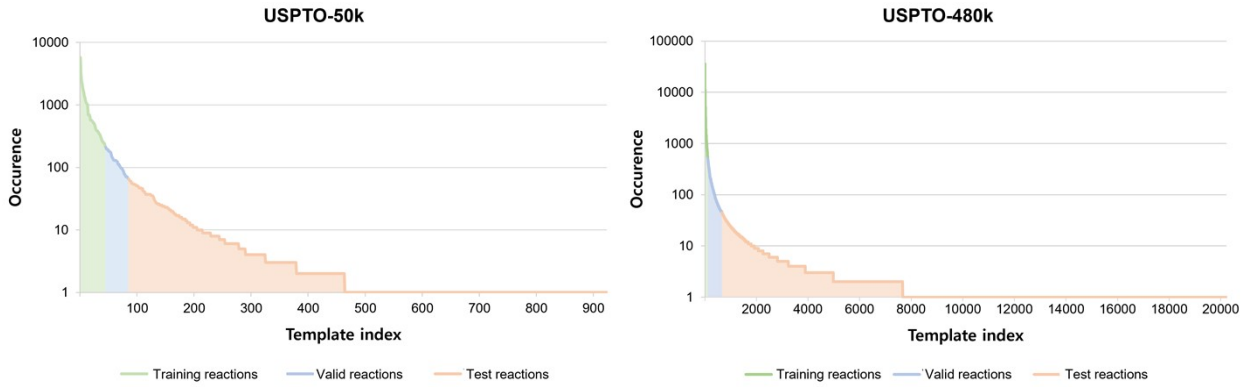

Figure S1. Distribution of dataset templates in the USPTO-50k and USPTO- 480k datasets, sorted by template popularity. The y-axis represents the number of occurrences for each template, and the x-axis ranks templates from most to least popular

S3. The performance of the evaluated surrogate models

Table S2. The exact match accuracy (%) of each reaction outcome prediction model trained with random split from USPTO-480k dataset. The highest values are highlighted in bold fonts.

Model \ Top-k	k=1	2	3	5
LocalTransform[1]	84.59	90.97	92.53	93.02
Chemformer[2]	84.79	90.66	92.19	93.05
Transformer[3]	84.58	90.83	92.54	93.49
MEGAN[4]	81.90	88.37	91.50	93.78

Table S3. The exact match accuracy (%) of each reaction outcome prediction model with template split (train and test reactions have different reaction templates) from USPTO-480k dataset. The highest values are highlighted in bold fonts.

Model \ Top-k	k=1	2	3	5
LocalTransform[1]	**8.68**	11.12	12.46	13.32
Chemformer[2]	7.96	**12.64**	**15.32**	**17.75**
Transformer[3]	8.03	11.17	13.08	15.20
MEGAN[4]	6.84	8.73	10.44	12.80

S4. The performance of the template-free models reported in the original reference

Table S4. The top-k exact match accuracy (%) reported in each reference on the USPTO-50k dataset for retrosynthesis prediction. The highest values are highlighted in bold fonts.

Model \ Top-k	k=1	3	5	10
Transformer[3]	-	-	-	-
Chemformer[2]	**54.3**	-	62.3	63.0
MEGAN[4]	48.1	**70.7**	**78.4**	**86.1**
GraphRetro[5]	53.6	**70.7**	74.6	77.0

S5. Round-trip evaluation of USPTO-50k dataset

Table S5: The top-k round-trip accuracy (%) for each evaluated model on the USPTO-50k and USPTO-480k datasets. The highest values are highlighted in bold fonts.

Model \ Top-k	USPTO-50k (%)			
	k=1	2	3	5
Transformer[3]	5.74	5.84	5.87	5.91
Chemformer[2]	10.54	12.35	13.51	14.86
MEGAN[4]	**48.97**	**48.10**	**48.05**	**48.01**
GraphRetro[5]	33.85	32.03	30.89	29.32

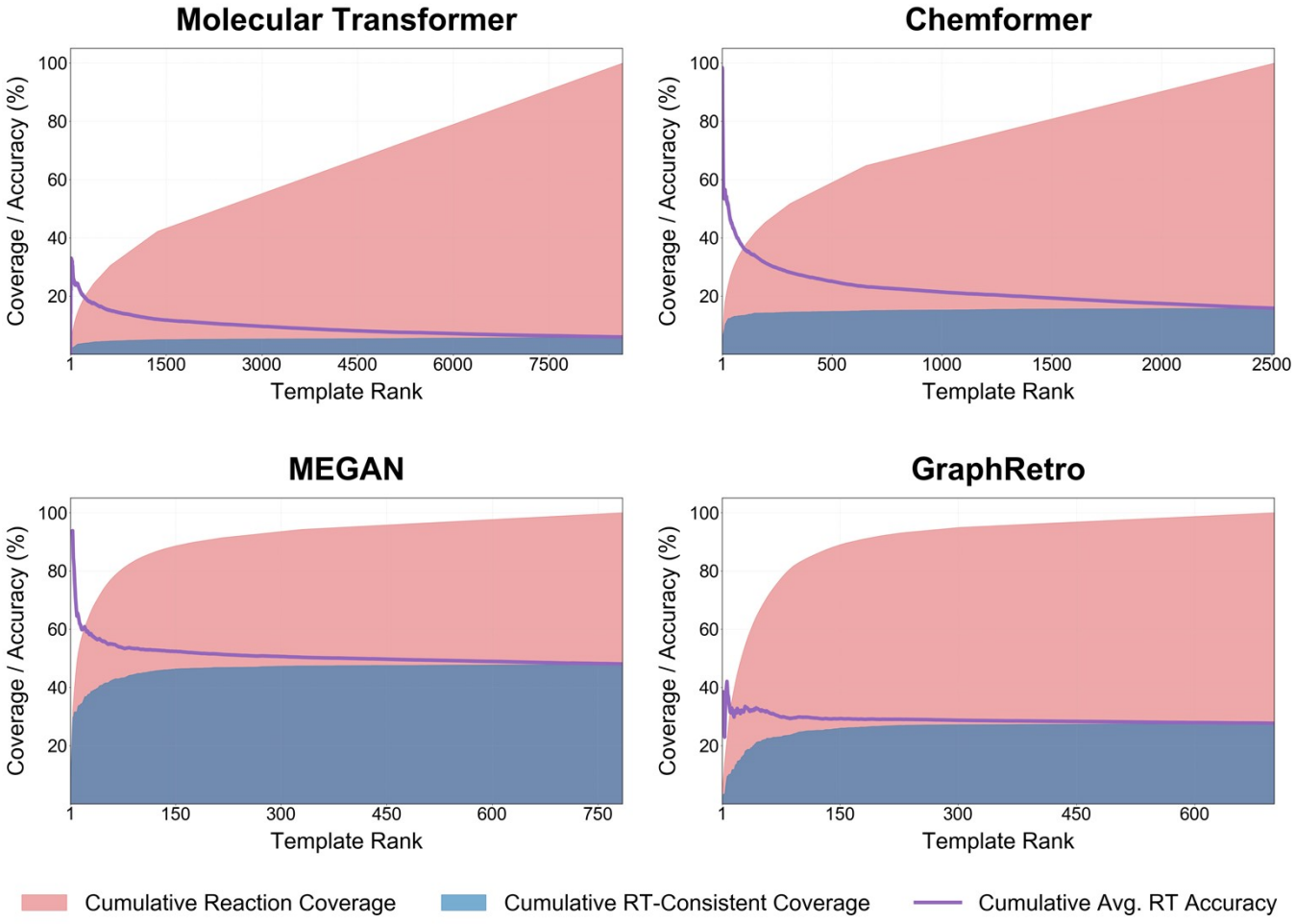

Figure S2. Cumulative reaction coverage, cumulative round-trip consistent coverage and cumulative average round-trip accuracy as a function of template rank of the novel reactions generated by the template-free models trained on USPTO-50k dataset.

S6. The statistics of template rank and round-trip accuracy

Table S6: The average round-trip accuracy (%) and reaction ratio for each evaluated model across different novel template rank ranges. The rank ranges are categorized based on the frequency of template usage in the USPTO-50k and USPTO-480k dataset. The values in parentheses indicate the ratio of reactions predicted by each model within the corresponding range. The highest accuracy values in each row are highlighted in bold fonts.

Dataset	Template Rank Range	Transformer[3]	Chemformer[2]	MEGAN[4]	GraphRetro[5]
USPTO-50k	Top 1-100	23.74% (13.8%)	36.24% (36.9%)	**53.05%** (84.5%)	29.75% (83.2%)
	Top 101-500	8.45% (13.8%)	6.27% (22.1%)	**23.26%** (11.9%)	18.79% (14.2%)
	Top 501-1,000	4.37% (8.9%)	3.67% (12.4%)	**10.88%** (3.6%)	8.96% (2.5%)
	Top 1,001-5,000	1.65% (34.6%)	**1.92%** (28.6%)	-	-
	Top >5,000	**1.64%** (29.0%)	-	-	-
USPTO-480k	Top 1-100	**32.18%** (7.8%)	28.69% (18.8%)	23.07% (36.9%)	25.55% (61.4%)
	Top 101-1,000	20.28% (15.7%)	13.16% (28.0%)	14.04% (37.9%)	**20.46%** (31.1%)
	Top 1,001-5,000	10.05% (19.4%)	5.70% (29.0%)	9.61% (23.4%)	**18.47%** (7.5%)
	Top 5,001-10,000	5.29% (11.4%)	3.86% (20.4%)	**8.41%** (1.7%)	-
	Top >10,000	3.02% (45.7%)	**3.76%** (3.8%)	-	-

Reference

(1) S. Chen and Y. Jung, Nat Mach Intell, 2022, 4, 772–780.
(2) R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, Mach. Learn.: Sci. Technol., 2022, 3, 015022.
(3) P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, ACS Cent. Sci., 2019, 5, 1572–1583.
(4) M. Sacha, M. Błaż, P. Byrski, P. Dąbrowski-Tumański, M. Chromiński, R. Loska, P. Włodarczyk-Pruszyński and S. Jastrzębski, J. Chem. Inf. Model., 2021, 61, 3273–3284.
(5) V. R. Somnath, C. Bunne, C. Coley, A. Krause and R. Barzilay, in Advances in Neural Information Processing Systems, 2021, 34, 9405–9415.