Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Models Matter: The Impact of Single-Step Retrosynthesis on Synthesis Planning†

Paula Torren-Peraire,*[a,b‡] Alan Kai Hassen,*[c,d‡] Samuel Genheden,[e] Jonas Verhoeven,[a] Djork-Arné Clevert,[d] Mike Preuss,[c] and Igor Tetko[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Retrosynthesis consists of breaking down a chemical compound recursively step-by-step into molecular precursors until a set of commercially available molecules is found with the goal to provide a synthesis route. Its two primary research directions, single-step retrosynthesis prediction, which models the chemical reaction logic, and multi-step synthesis planning, which tries to find the correct sequence of reactions, are inherently intertwined. Still, this connection is not reflected in contemporary research. In this work, we combine these two major research directions by applying multiple single-step retrosynthesis models within multi-step synthesis planning and analyzing their impact using public and proprietary reaction data. We find a disconnection between high single-step performance and potential route-finding success, suggesting that single-step models must be evaluated within synthesis planning in the future. Furthermore, we show that the commonly used single-step retrosynthesis benchmark dataset USPTO-50k is insufficient as this evaluation task does not represent model scalability or performance on larger and more diverse datasets. For multi-step synthesis planning, we show that the choice of the single-step model can improve the overall success rate of synthesis planning by up to +28% compared to the commonly used baseline model. Finally, we show that each single-step model finds unique synthesis routes, and differs in aspects such as route-finding success, the number of found synthesis routes, and chemical validity.

[a] *Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Neuherberg, Germany*
[b] *In-Silico Discovery, Janssen Research & Development, Janssen Pharmaceutica N.V, Beerse, Belgium*
[c] *Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands*
[d] *Machine Learning Research, Pfizer Worldwide Research Development and Medical, Berlin, Germany*
[e] *Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden*

∗ Corresponding Authors
‡ These authors contributed equally to this work

Supplementary Information

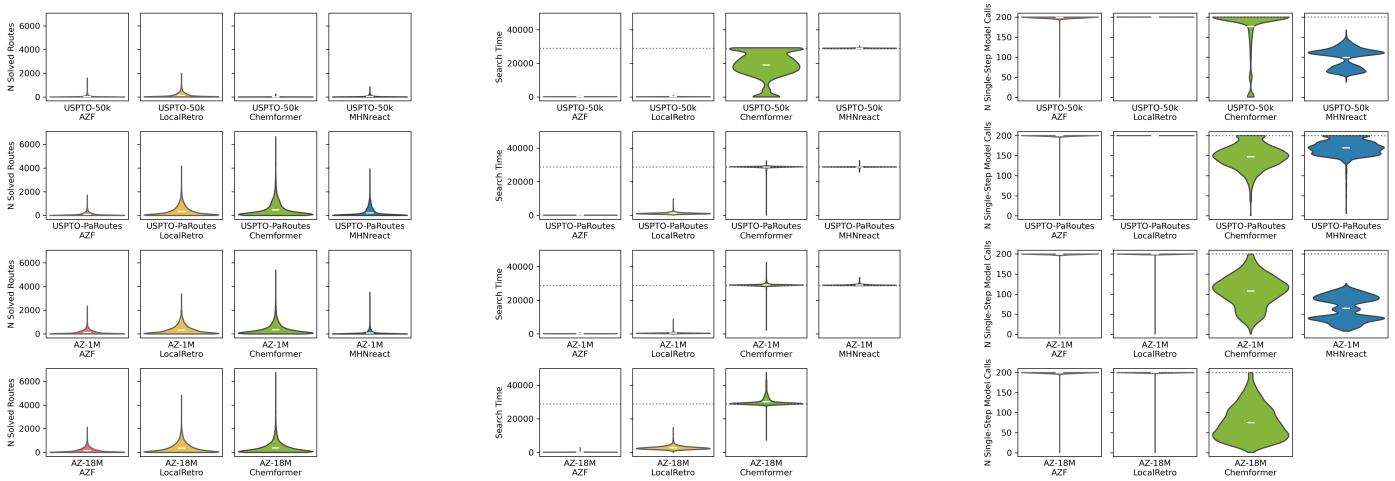
S1 Single-step retrosynthesis prediction

Table S1 Single-step Retrosynthesis Prediction Top-n Accuracy for AZF, LocalRetro, Chemformer, and MHNreact on the respective test sets of a dataset (USPTO-50k, USPTO-PaRoutes-1M, AZ-1M, AZ-18M)

		Top-N Accuracy (%)				
Training Dataset	**Model**	**Top-1**	**Top-3**	**Top-5**	**Top-10**	**Top-50**
USPTO-50k	AZF	41.6	62.5	69.5	75.8	77.4
	LocalRetro	52.0	76.5	84.6	90.7	96.2
	Chemformer	53.9	66.9	69.7	71.3	73.8
	MHNreact	49.4	73.8	81.1	87.3	93.1
USPTO-PaRoutes-1M	AZF	54.7	71.6	79.9	88.2	93.5
	LocalRetro	56.0	73.7	82.1	89.9	97.0
	Chemformer	54.8	74.6	80.6	86.0	92.6
	MHNreact	54.7	74.0	79.5	85.3	94.5
AZ-1M	AZF	19.9	28.6	33.0	38.5	42.3
	LocalRetro	24.4	34.5	39.2	44.9	53.6
	Chemformer	25.1	37.3	42.0	47.5	57.1
	MHNreact	22.3	32.1	35.8	40.2	49.2
AZ-18M	AZF	29.5	38.7	42.9	48.0	51.8
	LocalRetro	28.0	38.6	43.2	48.4	55.8
	Chemformer	45.0	62.6	68.5	74.5	83.1
	MHNreact	-	-	-	-	-

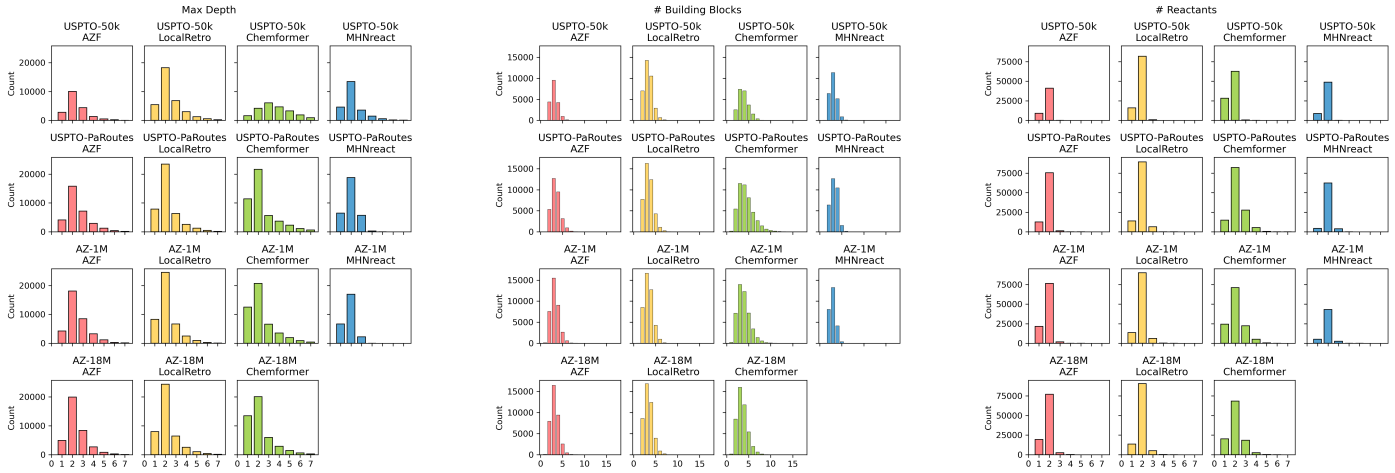
S2 Multi-step synthesis planning

S2.1 Caspyrus10k



(a) Solved routes (b) Search times (c) Single-step model calls

Fig. S1 Distributions of solved routes (a), search time (b) and single-step model calls (c) for synthesis planning results for all training datasets evaluated on Caspyrus10k. The dashed line indicates the respective limits set in algorithm search settings. The white line indicates the mean across all molecules for the shown model-training set combination.



(a) Maximum depth of routes (b) Building blocks per route (c) Number of reactants per reaction

Fig. S2 Statistics of top-5 found synthesis routes on Caspyrus10k by different single-step retrosynthesis models for all datasets. Shown are the maximum depth (a), referring to the longest linear path within the route, the number of building blocks within the route (b), and the number of reactants per route reaction (c)

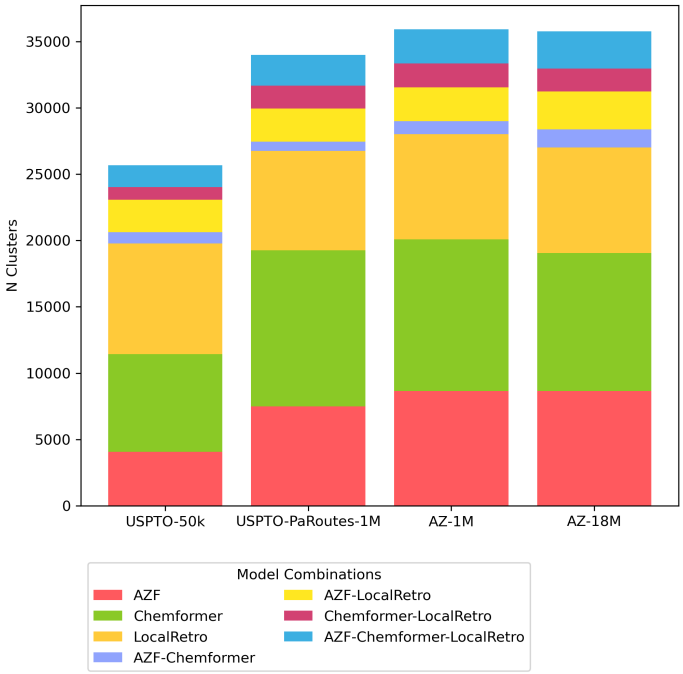


Fig. S3 Distribution and overlap of route clusters per single-step model (excluding MHNreact) and dataset when clustering with route-distance package[??]. Clusters were calculated on a per molecule basis, N clusters shows the number of clusters which contained the stated combination of models.

S2.2 Caspyrus10k Subsampling

Table S2 Multi-step synthesis planning metrics for a subsample size of 100 Caspyrus10k molecules. The performance is measured for each single-step model and dataset by randomly subsampling 1000 times with the subsample size (sampling without replacement). For each subsample, the same molecules are used across single-step models and datasets

		Overall	Average per Molecule		
Training Dataset	**Model**	**Success Rate (%)**	**Solved Routes**	**Search Time (s)**	**Model Calls**
USPTO-50k	AZF	41.2 ± 5.0	36.6 ± 8.6	159 ± 2	198 ± 1
	LocalRetro	74.2 ± 4.3	124 ± 18	160 ± 5	200 ± 0
	Chemformer	62.5 ± 4.7	7.28 ± 1.51	19043 ± 789	176 ± 5
	MHNreact	51.0 ± 5.0	38.8 ± 7.9	28956 ± 10	99.4 ± 2.4
USPTO-PaRoutes-1M	AZF	66.6 ± 4.8	84.2 ± 13.1	162 ± 1	199 ± 0
	LocalRetro	86.3 ± 3.3	326 ± 42	1217 ± 49	200 ± 0
	Chemformer	94.2 ± 2.4	464 ± 60	28811 ± 95	147 ± 2
	MHNreact	64.9 ± 4.7	215 ± 36	28839 ± 24	169 ± 1
AZ-1M	AZF	73.7 ± 4.4	124 ± 17	168 ± 4	199 ± 0
	LocalRetro	88.2 ± 3.1	322 ± 38	464 ± 34	199 ± 0
	Chemformer	94.6 ± 2.3	360 ± 44	29110 ± 68	107 ± 3
	MHNreact	56.0 ± 5.1	77.2 ± 16.9	29114 ± 33	64.6 ± 3.0
AZ-18M	AZF	76.4 ± 4.1	154 ± 21	153 ± 4	199 ± 1
	LocalRetro	87.4 ± 3.2	352 ± 43	2735 ± 109	199 ± 00
	Chemformer	91.0 ± 2.9	381 ± 50	30209 ± 242	75.2 ± 4.2

Table S3 Multi-step synthesis planning metrics for a subsample size of 500 Caspyrus10k molecules. The performance is measured for each single-step model and dataset by randomly subsampling 1000 times with the subsample size (sampling without replacement). For each subsample, the same molecules are used across single-step models and datasets

		Overall	Average per Molecule		
Training Dataset	**Model**	**Success Rate (%)**	**Solved Routes**	**Search Time (s)**	**Model Calls**
USPTO-50k	AZF	41.1 ± 2.1	36.2 ± 3.7	159 ± 0	198 ± 0
	LocalRetro	74.1 ± 1.8	124 ± 7	160 ± 2	200 ± 0
	Chemformer	62.5 ± 2.1	7.38 ± 0.66	19028 ± 337	176 ± 2
	MHNreact	51.1 ± 2.2	38.6 ± 3.4	28956 ± 4	99.4 ± 1.1
USPTO-PaRoutes-1M	AZF	66.4 ± 2.0	83.6 ± 5.8	162 ± 0	199 ± 0
	LocalRetro	86.1 ± 1.5	325 ± 18	1216 ± 21	200 ± 0
	Chemformer	94.2 ± 1.0	463 ± 26	28811 ± 41	147 ± 1
	MHNreact	64.7 ± 2.1	215 ± 15	28838 ± 10	169 ± 0
AZ-1M	AZF	73.7 ± 1.9	124 ± 7	168 ± 1	199 ± 0
	LocalRetro	88.1 ± 1.4	322 ± 16	464 ± 15	199 ± 0
	Chemformer	94.5 ± 1.0	358 ± 19	29108 ± 29	107 ± 1
	MHNreact	56.0 ± 2.2	77.2 ± 7.1	29116 ± 15	64.6 ± 1.4
AZ-18M	AZF	76.4 ± 1.8	154 ± 9	153 ± 2	199 ± 0
	LocalRetro	87.3 ± 1.4	351 ± 19	2732 ± 48	199 ± 0
	Chemformer	91.0 ± 1.2	380 ± 22	30212 ± 110	75.1 ± 1.8

Table S4 Multi-step synthesis planning metrics for a subsample size of 1,000 Caspyrus10k molecules. The performance is measured for each single-step model and dataset by randomly subsampling 1000 times with the subsample size (sampling without replacement). For each subsample, the same molecules are used across single-step models and datasets.

		Overall	Average per Molecule		
Training Dataset	**Model**	**Success Rate (%)**	**Solved Routes**	**Search Time (s)**	**Model Calls**
USPTO-50k	AZF	41.1 ± 1.4	36.1 ± 2.6	159 ± 0	198 ± 0
	LocalRetro	74.0 ± 1.3	124 ± 5	160 ± 1	200 ± 0
	Chemformer	62.4 ± 1.4	7.35 ± 0.47	19061 ± 245	176 ± 1
	MHNreact	50.9 ± 1.5	38.5 ± 2.3	28956 ± 3	99.4 ± 0.7
USPTO-PaRoutes-1M	AZF	66.3 ± 1.5	83.5 ± 4.1	162 ± 0	199 ± 0
	LocalRetro	86.0 ± 1.1	324 ± 13	1218 ± 15	200 ± 0
	Chemformer	94.1 ± 0.7	463 ± 18	28811 ± 29	147 ± 0
	MHNreact	64.6 ± 1.5	214 ± 11	28839 ± 7	169 ± 0
AZ-1M	AZF	73.5 ± 1.4	124 ± 5	168 ± 1	199 ± 0
	LocalRetro	88.0 ± 1.0	321 ± 11	465 ± 10	199 ± 0
	Chemformer	94.4 ± 0.7	358 ± 13	29108 ± 20	107 ± 1
	MHNreact	56.0 ± 1.5	76.9 ± 5.1	29115 ± 10	64.6 ± 0.9
AZ-18M	AZF	76.2 ± 1.3	154 ± 6	153 ± 1	199 ± 0
	LocalRetro	87.3 ± 1.0	350 ± 13	2737 ± 33	199 ± 0
	Chemformer	90.9 ± 0.9	381 ± 14	30210 ± 79	75.1 ± 1.3

Table S5 Multi-step synthesis planning metrics for a subsample size of 5,000 Caspyrus10k molecules. The performance is measured for each single-step model and dataset by randomly subsampling 1000 times with the subsample size (sampling without replacement). For each subsample, the same molecules are used across single-step models and datasets

		Overall	Average per Molecule		
Training Dataset	**Model**	**Success Rate (%)**	**Solved Routes**	**Search Time (s)**	**Model Calls**
USPTO-50k	AZF	41.1 ± 0.5	36.0 ± 0.9	159 ± 0	198 ± 0
	LocalRetro	74.0 ± 0.4	124 ± 1	160 ± 0	200 ± 0
	Chemformer	62.3 ± 0.5	7.37 ± 0.16	19053 ± 79	176 ± 0
	MHNreact	50.9 ± 0.5	38.4 ± 0.8	28956 ± 1	99.3 ± 0.2
USPTO-PaRoutes-1M	AZF	66.3 ± 0.5	83.4 ± 1.4	162 ± 0	199 ± 0
	LocalRetro	86.0 ± 0.3	324 ± 4	1218 ± 4	200 ± 0
	Chemformer	94.1 ± 0.2	463 ± 6	28810 ± 10	147 ± 0
	MHNreact	64.6 ± 0.5	214 ± 3	28838 ± 2	169 ± 0
AZ-1M	AZF	73.5 ± 0.4	124 ± 1	168 ± 0	199 ± 0
	LocalRetro	88.0 ± 0.3	321 ± 3	465 ± 3	199 ± 0
	Chemformer	94.4 ± 0.2	358 ± 4	29109 ± 7	107 ± 0
	MHNreact	55.9 ± 0.5	77.0 ± 1.7	29115 ± 3	64.6 ± 0.3
AZ-18M	AZF	76.2 ± 0.4	154 ± 2	153 ± 0	199 ± 0
	LocalRetro	87.3 ± 0.3	350 ± 4	2737 ± 10	199 ± 0
	Chemformer	90.9 ± 0.3	380 ± 4	30212 ± 26	75.1 ± 0.4

Table S6 Multi-step synthesis planning metrics for the provided randomly selected subsample of 1,000 Caspyrus10k molecules

		Overall	Average per Molecule		
Training Dataset	**Model**	**Success Rate (%)**	**Solved Routes**	**Search Time (s)**	**Model Calls**
USPTO-50k	AZF	40.7	37.5	159	198
	LocalRetro	73.6	125	163	200
	Chemformer	62.9	7.09	19269	176
	MHNreact	50.0	39.3	28955	99.0
USPTO-PaRoutes-1M	AZF	67.4	87.7	162	199
	LocalRetro	85.6	327	1231	200
	Chemformer	94.1	448	28752	146
	MHNreact	63.6	204	28845	168
AZ-1M	AZF	74.8	126	169	199
	LocalRetro	88.6	325	466	200
	Chemformer	94.2	382	29087	108
	MHNreact	54.5	75.8	29113	65.1
AZ-18M	AZF	77.4	156	154	199
	LocalRetro	87.5	351	2783	200
	Chemformer	90.7	391	30127	76.1

S2.3 PaRoutes

Table S7 Multi-step synthesis planning route accuracy (a) and building block accuracy (b) on PaRoutes gold-standard synthesis routes with different single-step models trained on USPTO-PaRoutes-1M

(a) Route Accuracy

Training Dataset	Model	Top-1	Top-3	Top-5	Top-10	Top-50
USPTO-PaRoutes-1M	AZF	23.7	48.5	56.5	60.7	61.8
	LocalRetro	3.72	9.92	13.8	20.2	36.0
	Chemformer	1.9	5.8	9.4	13.8	26.5
	MHNreact	4.2	11.3	16.0	22.9	39.7

(b) Building Block Accuracy

Training Dataset	Model	Top-1	Top-3	Top-5	Top-10	Top-50
USPTO-PaRoutes-1M	AZF	45.3	64.1	71.2	75.2	76.0
	LocalRetro	16.4	28.3	34.7	43.8	62.6
	Chemformer	9.8	20.3	25.8	33.9	49.5
	MHNreact	15.6	26.9	33.2	41.3	57.1

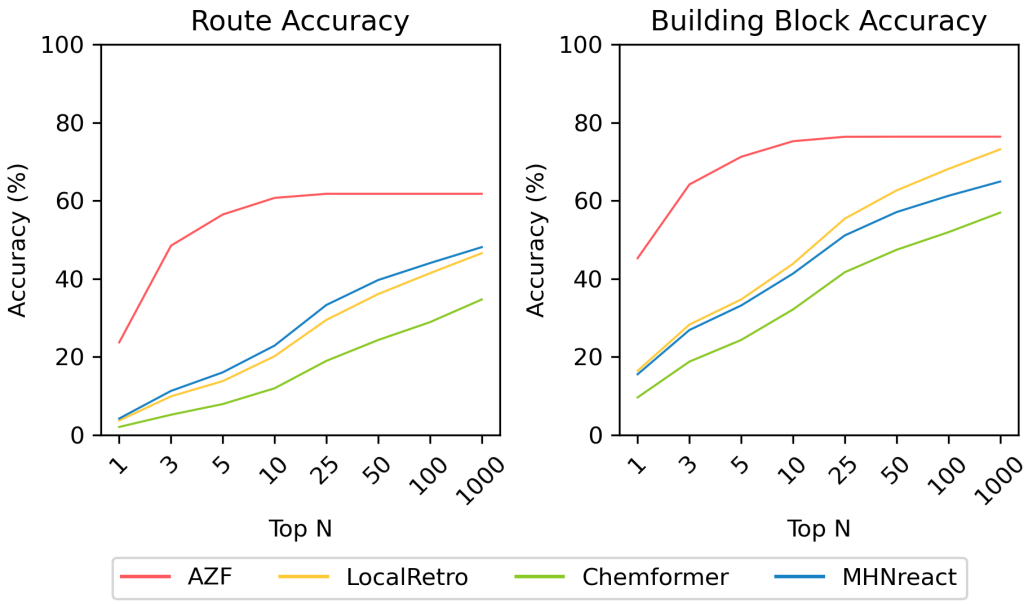


Fig. S4 Multi-step synthesis planning accuracy up to top-1000 on PaRoutes gold-standard synthesis routes with different single-step models trained on USPTO-PaRoutes-1M. Route accuracy measures the ability to recover the correct synthesis route within top-n, whereas building block accuracy measures the ability to recover the correct building blocks while not considering reactions and intermediates.

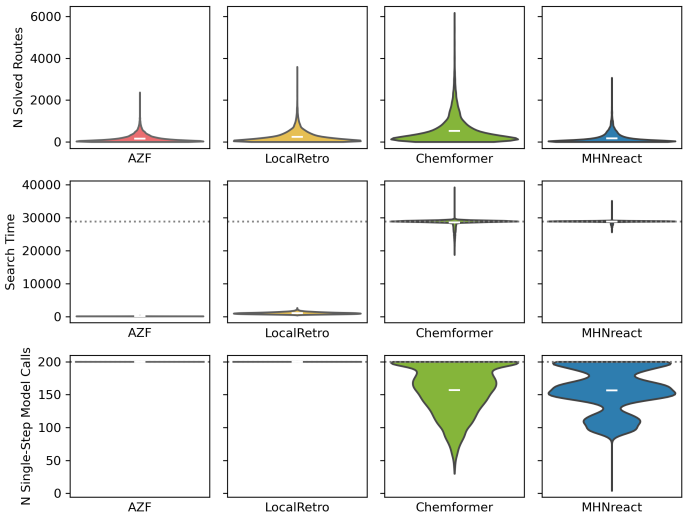


Fig. S5 Distributions of solved routes, search time and single-step model calls for synthesis planning results of single-step models trained on USPTO-PaRoutes-1M and evaluated on PaRoutes. The dashed line indicates the respective limits set in algorithm search settings. The white line indicates the mean across all molecules for the shown model-training set combination.

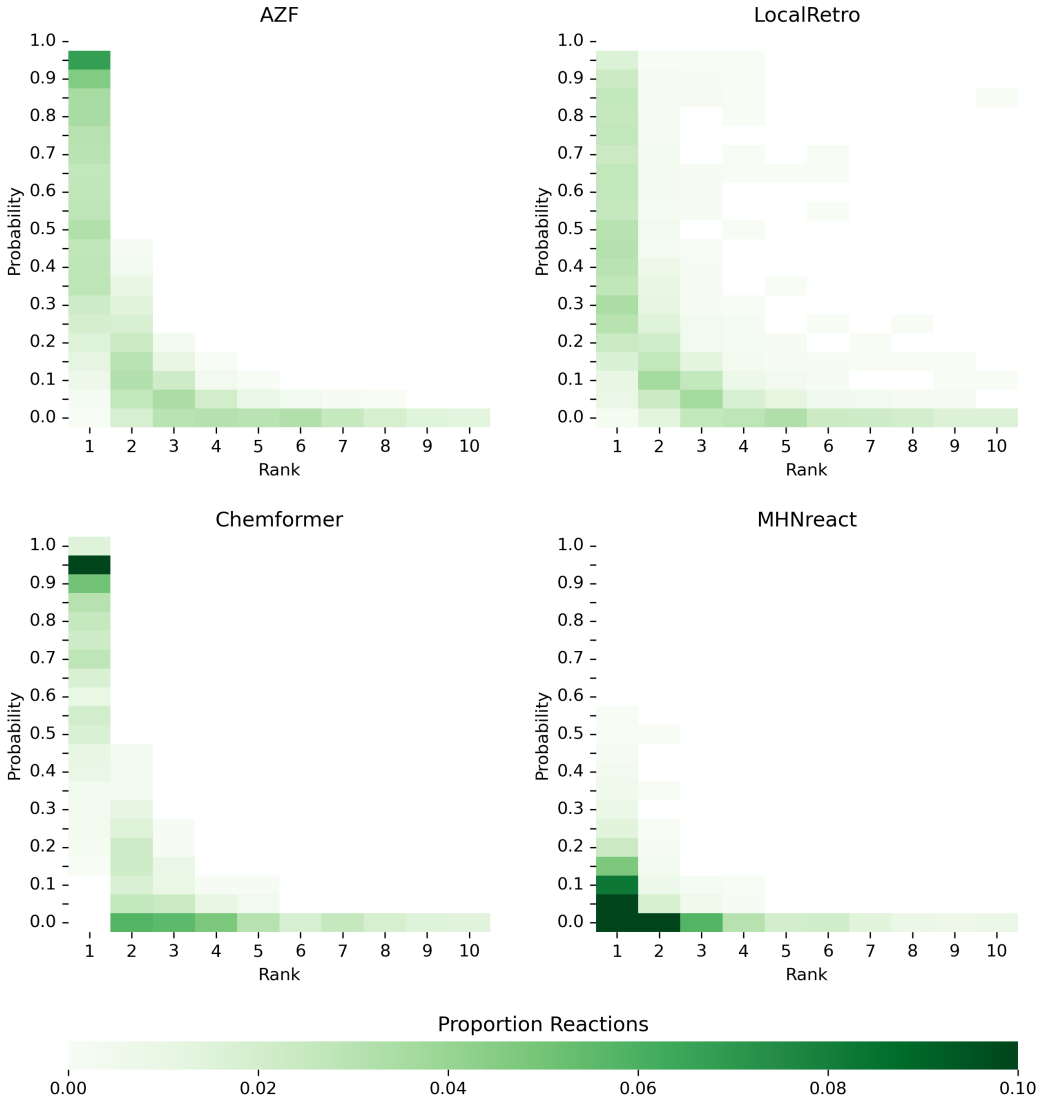


Fig. S6 Single-step model prior and rank distributions of reactions from the correctly predicted PaRoutes synthesis routes. Reactions are extracted from the top-10 predicted routes for each single-step retrosynthesis model trained on USPTO-PaRoutes-1M.