Supporting information for

A Simple Similarity Metric for Comparing Synthetic Routes

Samuel Genheden and Jason D. Shields

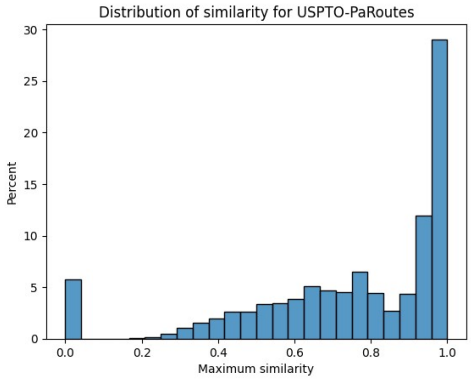
Contents

Note that atom mapping for the atorvastatin and strychnine case studies are provided as zipped png files, as they are too unwieldy to view easily in a Word document. In these examples, some reactants are abbreviated to provide only their reaction centers, e.g. a cobalt-mediated [2+2+2] cyclization in Vollhardt's strychnine synthesis uses $CpCo(C_2H_4)_2$ as the cobalt reagent. This has been abbreviated as "Co" in the reaction data file and corresponding png. Similarly, in the biosynthesis O_2 is used as the oxidant even though the true oxidant is likely much more complex. These simplifications made it easier to catch errors and they have no effect on the score.

Additional details and results on retrosynthesis benchmarking

We performed retrosynthesis experiments on three datasets of routes. First, we used the PaRoutes *n*1-set, previously described.[1] This set where extracted from reaction data from the US Patent and Trademark office, by grouping reactions originating from the same patent. The *n*1-set consists of 10,000 such routes selected based on route diversity. Second, we use a selection 4934 routes from Journal of Medicinal Chemistry (JMC). These were selected from the full set of JMC routes previously described by selecting the longest route from each unique publication. Third, we extracted a new set of routes from USPTO that we will call USPTO MedChem. USPTO MedChem reference routes were extracted from the USTPO dataset as previously described. The targets were cross-checked with routes extracted from Journal of Medicinal Chemistry to identify relevant targets. The USPTO MedChem routes are provided on Zenodo for download.

All retrosynthesis experiments were carried out using version 4.0 of AiZynthFinder. We employed a filter model trained on USPTO data together with the expansion policy (one-step retrosynthesis model). We performed 500 iterations of Monte Carlo Tree Search, without any time limit. The maximum depth was set to 10 for the PaRoutes targets, and 7 for the JMC and USTPO MedChem targets. For the PaRoutes experiments a stock consisting of the starting material in the reference routes where used. For the JMC and USPTO MedChem, we also used a stock consisting of the starting material in the reference routes, but also a stock consisting of molecules downloaded from the eMolecules website in January 2023.



To complement the results in Table 2, we performed the retrosynthesis predictions with the full eMolecules stock (Table S1). Similar conclusions can be drawn from these experiments, although using this expanded set of possible starting materials increases the success rate while decreasing the top-N accuracies.

Table S1 - Benchmarking of three retrosynthesis model on a dataset of 4934 targets and routes from Journal of Medicinal Chemistry using full eMolecules stock

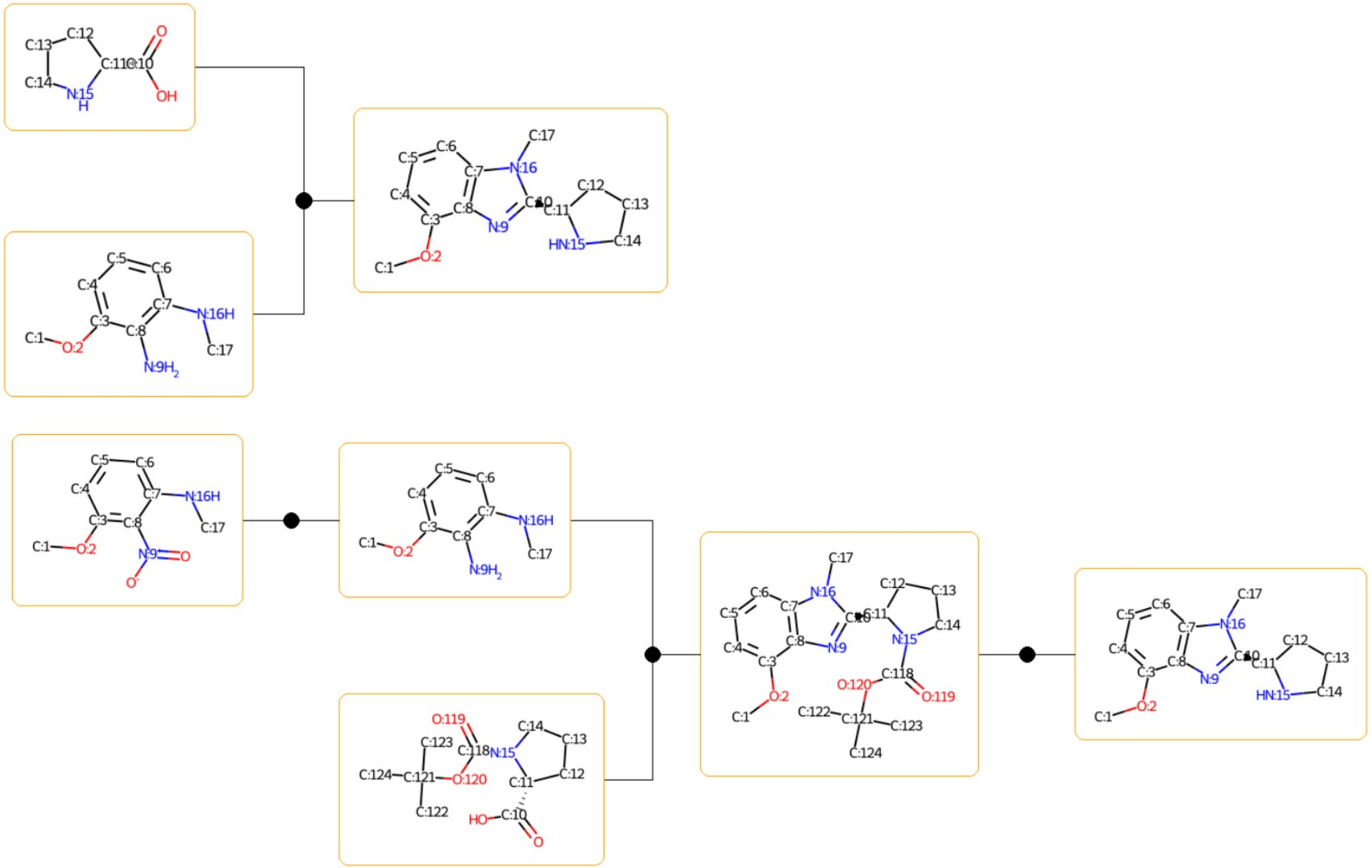
Model	Success rate	Accuracy		Similarity
		top-1	top-10	
USPTO-PaRoutes[1]	86.6%	0.02	0.08	0.75
AZ-2019	87.9%	0.01	0.08	0.70
AZ-retrained[2]	95.1%	0.02	0.12	0.82

Finally, we constructed a new set of reference routes from the USPTO dataset by identifying 667 of these targets for which a route has been published in Journal of Medicinal Chemistry (see Table S2). This constitutes a completely open set of routes that are still relevant for pharmaceutical applications. When comparing AZ-Retrained to AZ-2019, we observe a smaller difference in all performance metrics than we did on the JMC routes, although AZ-Retrained still outperforms AZ-2019 (especially when employing only reference stock). However, in this case there is only a 0.01 average difference between AZ-Retrained and USPTO-PaRoutes, indicating essentially equivalent performance between these two models when using this dataset.

Table S2 - Benchmarking of three retrosynthesis model on a dataset of medicinal chemistry targets from USPTO

Model	Success rate	Accuracy		Similarity
		top-1	top-10	
	Stock: eMolecules			
USPTO-PaRoutes[1]	96.7%	0.06	0.17	0.84
AZ-2019[3]	96.9%	0.04	0.16	0.83
AZ-retrained[2]	99.6%	0.04	0.16	0.86
	Stock: reference routes			
USPTO-PaRoutes[1]	93.3%	0.47	0.74	0.93
AZ-2019	88.7%	0.28	0.64	0.88
AZ-retrained[2]	96.0%	0.33	0.68	0.94

Atom mapping and demo calculation for example described in Figure 1



Calculation of S_{atom}:

$m_{1,1}$ = {10,11,12,13,14,15} – this is proline

$m_{1,2}$ = {1,2,3,4,5,6,7,8,9,16,17} – this is the phenylenediamine

$m_{2,1}$ = {1,2,3,4,5,6,7,8,9,16,17} – this is the nitroaniline

$m_{2,2}$ = {1,2,3,4,5,6,7,8,9,16,17} – this is the phenylendiamine that came from reducing the nitroaniline

$m_{2,3}$ = {10,11,12,13,14,15} – this is the boc-protected proline. Note that the boc group comprising atoms 118, 119, 120, 121, 122, 123, and 124 are not included as these atoms are not part of the final compound.

$m_{2,4}$ = {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17} – this is the boc-protected penultimate compound in route 2

$O(m_{1,1},m_{2,1})$ = |{10,11,12,13,14,15} ∩ {1,2,3,4,5,6,7,8,9,16,17} |/ max(|{10,11,12,13,14,15}|,|{1,2,3,4,5,6,7,8,9,16,17}|)

= 0/11 = 0

$O(m_{1,1},m_{2,2})$ = |{10,11,12,13,14,15} ∩ {1,2,3,4,5,6,7,8,9,16,17} |/ max(|{10,11,12,13,14,15}|,|{1,2,3,4,5,6,7,8,9,16,17}|)

= 0/11 = 0

$O(m_{1,1},m_{2,3})$ = |{10,11,12,13,14,15} ∩ {10,11,12,13,14,15} |/ max(|{10,11,12,13,14,15}|,|{10,11,12,13,14,15}|)

= 6/6 = 1

$O(m_{1,1},m_{2,4})$ = |{10,11,12,13,14,15} ∩ {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17}| / max(|{10,11,12,13,14,15}|,|{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17}|)

= 6/17 = 0.353

$O(m_{1,2},m_{2,1})$ = 1

$O(m_{1,2},m_{2,2})$ = 1

$O(m_{1,2},m_{2,3})$ = 0

$O(m_{1,2},m_{2,4})$ = |{1,2,3,4,5,6,7,8,9,16,17} ∩ {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17} |/
max(|{1,2,3,4,5,6,7,8,9,16,17}|,|{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17}|)

= 11/17 = 0.647

$O(m_{2,1},m_{1,1}) = O(m_{1,1},m_{2,1}) = 0$

$O(m_{2,1},m_{1,2}) = O(m_{1,2},m_{2,1}) = 1$

$O(m_{2,2},m_{1,1}) = O(m_{1,1},m_{2,2}) = 0$

$O(m_{2,2},m_{1,2}) = O(m_{1,2},m_{2,2}) = 1$

$O(m_{2,3},m_{1,1}) = O(m_{1,1},m_{2,3}) = 1$

$O(m_{2,3},m_{1,2}) = O(m_{1,2},m_{2,3}) = 0$

$O(m_{2,4},m_{1,1}) = O(m_{1,1},m_{2,4}) = 0.353$

$O(m_{2,4},m_{1,2}) = O(m_{1,1},m_{2,4}) = 0.647$

$S_{atom} = ((1 + 1) + (1 + 1 + 1 + 0.647))/(2 + 4) = 5.647/6 = 0.942$

Calculation of S_{bond}:

$r_{1,1} = \{(9,10), (10,16)\}$ – cyclization reaction

$r_{2,1} = \{\}$ – here the nitro group is reduced which does not form any bonds between heavy atoms

$r_{2,2} = \{(9,10), (10,16)\}$ – cyclization reaction

$r_{2,3} = \{\}$ – here the boc group is removed which is the same situation as above with nitro reduction

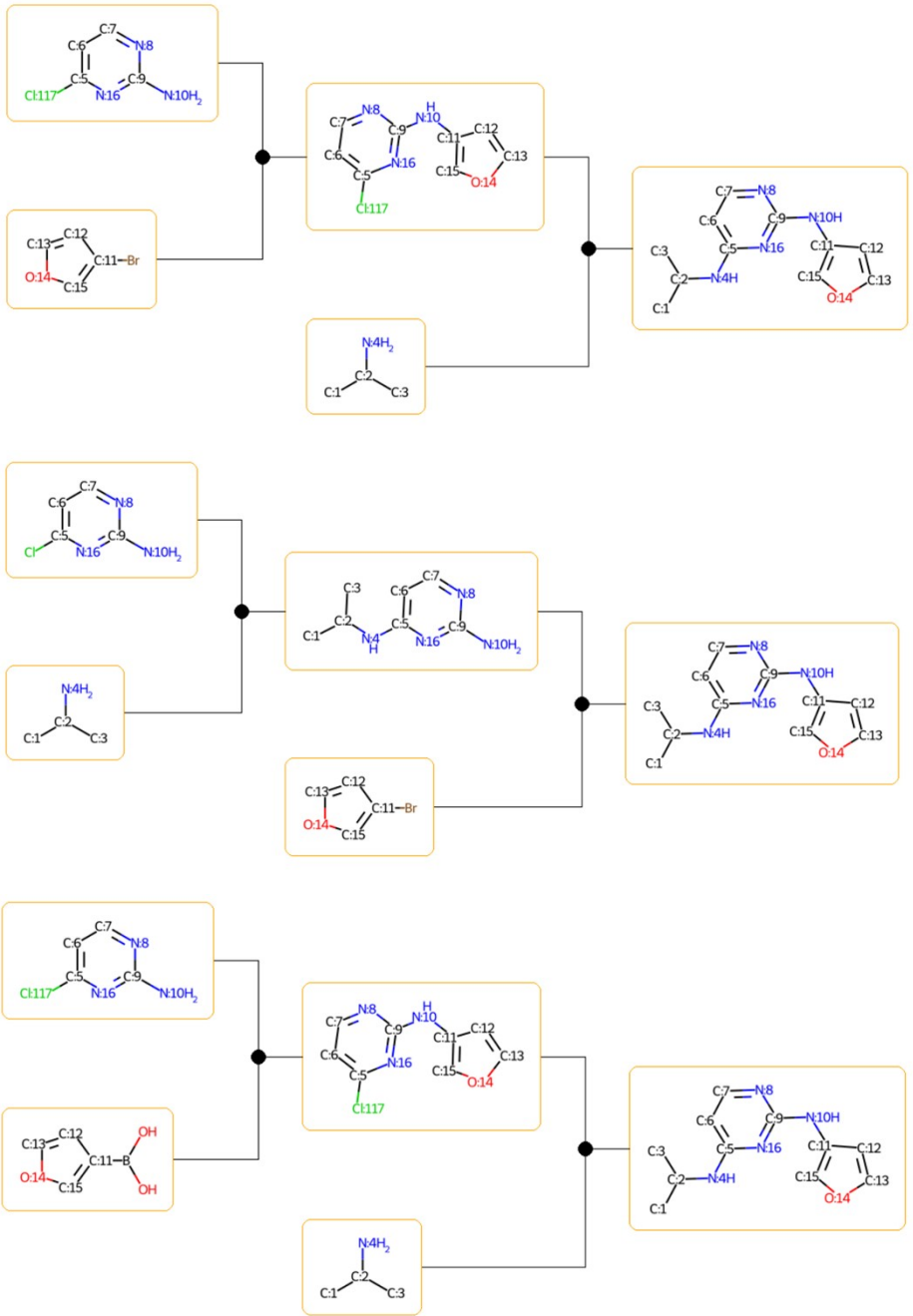
$\rho_1 = \{(9,10), (10,16)\}$

$\rho_2 = \{(9,10), (10,16)\}$

$S_{bond} = |\{(9,10),(10,16)\} \cap \{(9,10),(10,16)\}|/\max(1,1) = 1$

$S_{1,2} = \text{G-mean}(0.942, 1) = \sqrt{(0.942 \times 1)} = 0.970$

Atom mapping for example described in Figure 2



Atorvastatin comparison table

Table S3. Pairwise comparison scores for the 4 atorvastatin syntheses in Figure 3. The upper part of the table shows the route similarity and the lower part of the table shows the TED.

	A	B	C	D
A		0.88	0.59	0.49
B	7.99		0.62	0.45
C	24.72	30.29		0.74
D	26.81	31.48		

Strychnine comparison table

Table S4. Pairwise comparison scores for the 12 strychnine syntheses. Reissig, Vollhardt, Woodward, and Rawal used isostrychnine as the penultimate intermediate.

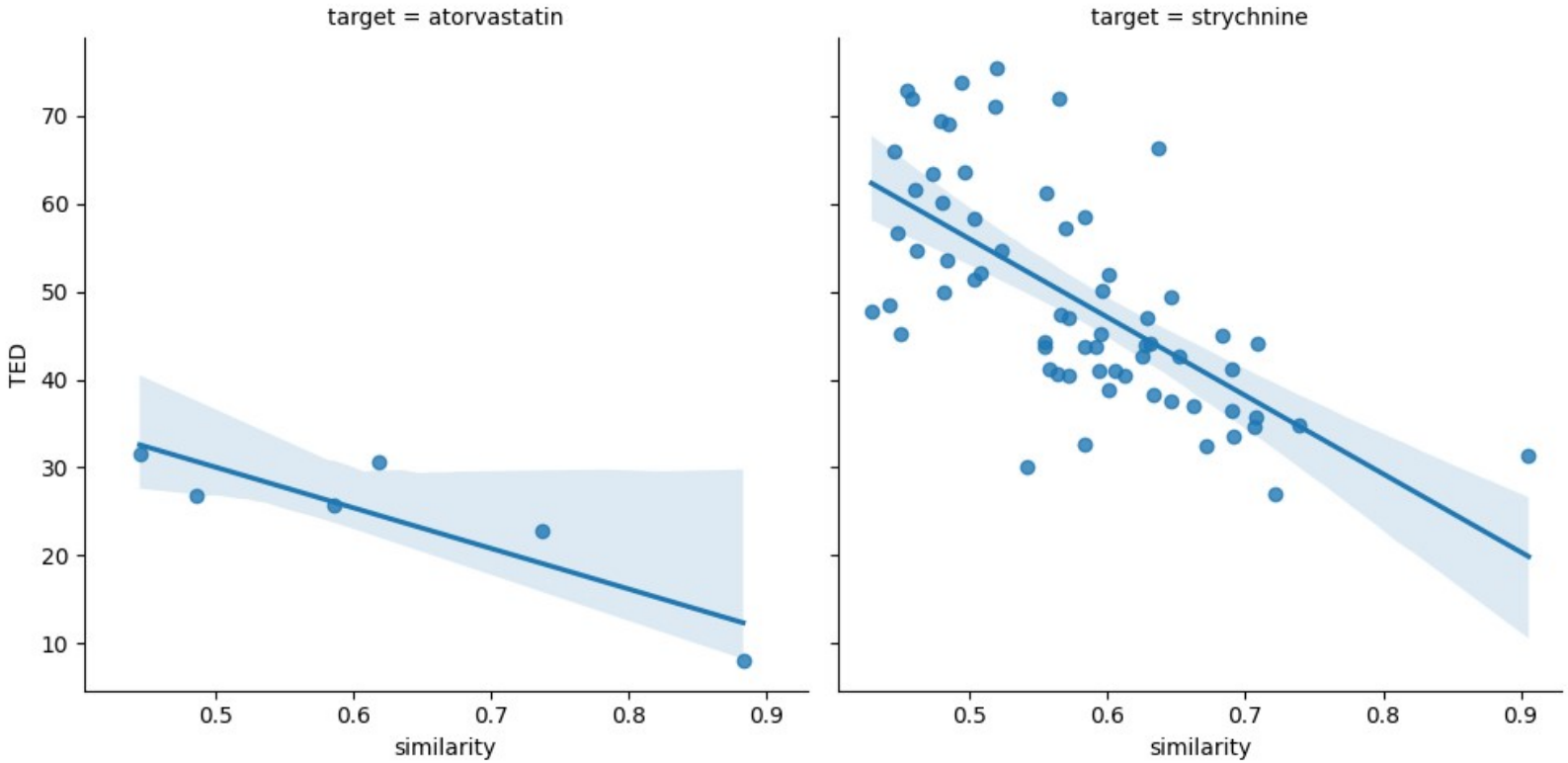
	Biosynthesis	Fukuyama	Kuehne enantioselective	Kuehne racemic	MacMillan	Martin	Overman	Reissig	Vollhardt	Woodward	Rawal	Vanderwal
Biosynthesis	1.00	0.59	0.69	0.67	0.56	0.74	0.51	0.61	0.59	0.44	0.43	0.69
Fukuyama	0.59	1.00	0.57	0.55	0.50	0.56	0.58	0.44	0.48	0.49	0.50	0.55
Kuehne ent.	0.69	0.57	1.00	0.91	0.65	0.65	0.48	0.60	0.60	0.47	0.45	0.68
Kuehne rac.	0.67	0.55	0.91	1.00	0.63	0.63	0.46	0.60	0.60	0.48	0.45	0.66
MacMillan	0.56	0.50	0.65	0.63	1.00	0.71	0.56	0.57	0.63	0.45	0.60	0.65
Martin	0.74	0.56	0.65	0.63	0.71	1.00	0.52	0.63	0.63	0.48	0.57	0.71
Overman	0.51	0.58	0.48	0.46	0.56	0.52	1.00	0.46	0.50	0.56	0.64	0.57
Reissig	0.61	0.44	0.60	0.60	0.57	0.63	0.46	1.00	0.72	0.48	0.54	0.71
Vollhardt	0.59	0.48	0.60	0.60	0.63	0.63	0.50	0.72	1.00	0.52	0.58	0.69
Woodward	0.44	0.49	0.47	0.48	0.45	0.48	0.56	0.48	0.52	1.00	0.52	0.46
Rawal	0.43	0.50	0.45	0.45	0.60	0.57	0.64	0.54	0.58	0.52	1.00	0.58
Vanderwal	0.69	0.55	0.68	0.66	0.65	0.71	0.57	0.71	0.69	0.46	0.58	1.00

Table S5. Pairwise TED (tree edit distances) for the 12 strychnine syntheses.

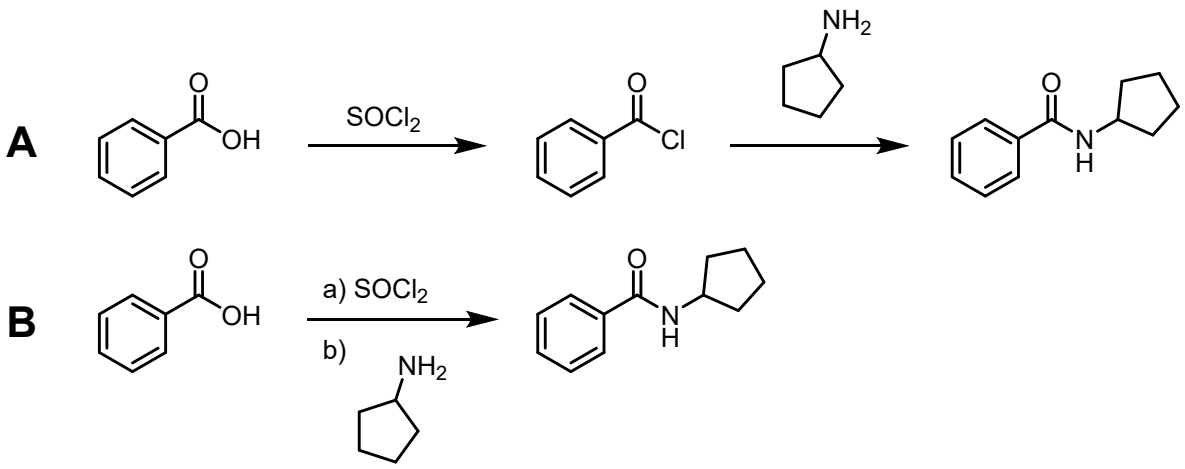
	Biosynth	Fukuyama	Kuehne ent	Kuehne rac	MacMillan	Martin	Overman	Rawal	Reissig	Vanderwal	Vollhardt	Woodward
Biosynth	0.00	40.97	36.47	32.49	41.21	34.77	52.21	47.69	40.39	33.57	43.74	65.92
Fukuyama	40.97	0.00	47.22	43.95	51.35	41.00	56.55	57.56	49.77	43.11	52.89	72.66
Kuehne ent	36.47	47.22	0.00	31.48	49.28	42.72	50.53	55.63	50.97	44.71	51.00	63.34
Kuehne rac	32.49	43.95	31.48	0.00	43.89	38.26	54.77	45.20	38.89	36.91	41.10	60.18
MacMillan	41.21	51.35	49.28	43.89	0.00	44.04	62.01	46.34	40.55	37.75	42.50	72.18
Martin	34.77	41.00	42.72	38.26	44.04	0.00	54.25	48.16	44.90	35.76	45.97	69.29
Overman	52.21	56.55	50.53	54.77	62.01	54.25	0.00	67.25	61.69	57.13	64.80	72.02
Rawal	47.69	57.56	55.63	45.20	46.34	48.16	67.25	0.00	30.02	43.88	31.92	75.22
Reissig	40.39	49.77	50.97	38.89	40.55	44.90	61.69	30.02	0.00	34.61	27.35	69.45
Vanderwal	33.57	43.11	44.71	36.91	37.75	35.76	57.13	43.88	34.61	0.00	41.02	71.79
Vollhardt	43.74	52.89	51.00	41.10	42.50	45.97	64.80	31.92	27.35	41.02	0.00	70.17
Wodward	65.92	72.66	63.34	60.18	72.18	69.29	72.02	75.22	69.45	71.79	70.17	0.00

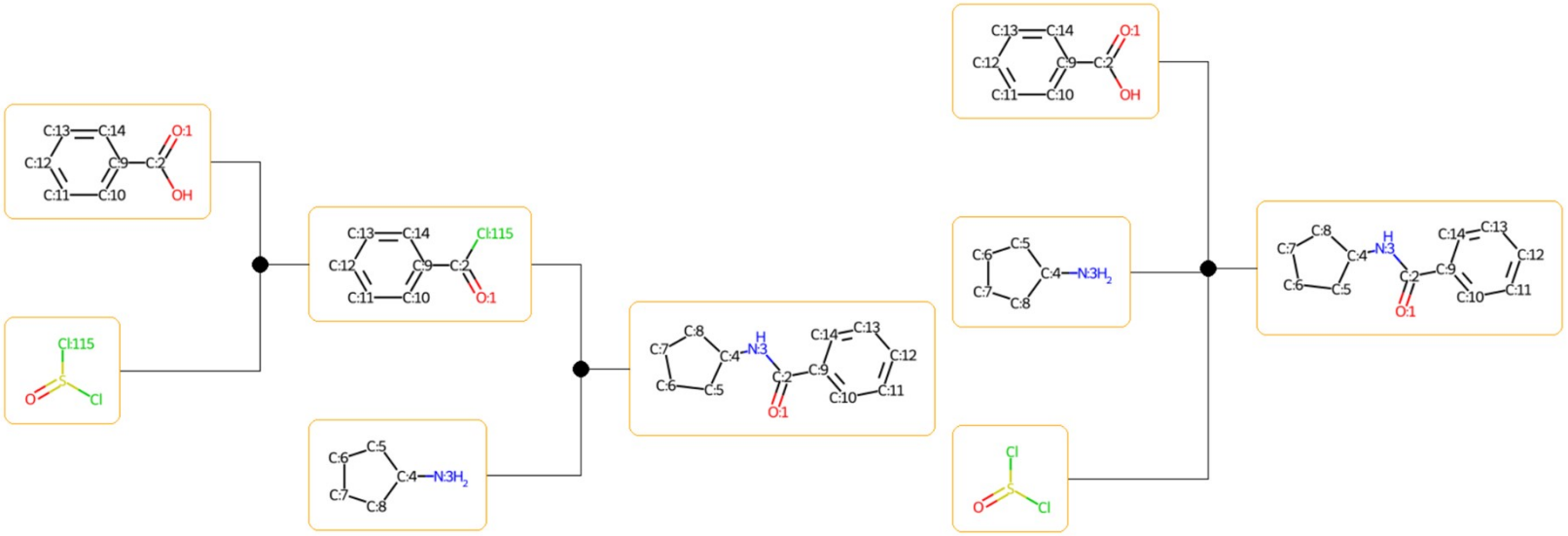
Correlation between route similarity and TED

The squared Pearson correlation coefficient, r^2 is 0.77 for the atorvastatin routes and 0.44 for the strychnine routes.



Effect of how a telescoped reaction is reported





$S_{A,B}$ = 0.845

References

1. S. Genheden and E. Bjerrum, *Digital Discovery*, 2022, **1**, 527-539.
2. S. Genheden, P.-O. Norrby and O. Engkvist, *Journal of Chemical Information and Modeling*, 2023, **63**, 1841-1846.
3. A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist and E. J. Bjerrum, *Chemical Science*, 2020, **11**, 154-168.