TotSyn: A Total Synthesis Reaction Dataset for Machine Learning in Organic Chemistry

Anonymous Authors[1]

Abstract

Total synthesis is the construction of complex natural products from simple, commercially available starting materials. Despite its importance in synthetic chemistry, it remains underrepresented in machine learning resources, as existing reaction datasets are largely derived from patent data such as USPTO and fail to capture its distinctive chemical complexity. Here, we introduce TotSyn, a curated total synthesis reaction dataset collected from peer-reviewed journals. We analyze its reaction template diversity and longest linear sequence (LLS) and compare them with USPTO-50k, revealing substantial differences in template distribution and route complexity. Our results highlight the limitations of patent-derived datasets and position TotSyn as a benchmark resource for machine learning on total synthesis.

1. Scientific bottleneck

Total synthesis is the complete laboratory construction of complex natural products from simpler, commercially available starting materials, serving as a cornerstone of organic chemistry that drives innovations in synthetic methods and reaction design. While prior works have focused on quantifying molecular complexity of natural products (Tyrin et al., 2025) or improving single-step and multi-step retrosynthesis performance trained on augmented datasets based on USPTO data (Sathyanarayana et al., 2025; Deng et al., 2025; Zheng et al., 2022), little attention has been paid to linking reaction template diversity to natural product synthesis. Furthermore, existing machine learning datasets primarily derived from USPTO may not adequately represent the chemical complexity inherent to total synthesis.

In this study, we construct a total synthesis dataset curated from published journals and provide a comprehensive analysis encompassing both reaction template diversity and the longest linear sequence (LLS), which represents the longest sequential chain of reactions required to complete a total synthesis. We further compare our dataset against USPTO-50k (Dai et al., 2019) to highlight key distributional differences in template usage and route complexity.

Table 1. Comparison of dataset scale and reaction template coverage across TotSyn, USPTO-50k, and PaRoutes.

Dataset	TotSyn	USPTO-50k	PaRoutes
Total reaction	105,002	50,000	347,040
Total template	30,703	2,367	10,601

2. Acquisition roadmap

Reactions were collected from Reaxys (Elsevier, 2026) by querying publications with "total synthesis" in the title. We restricted the search to three major organic chemistry journals—*Organic Letters* (1,573 papers), *JACS* (1,939 papers), and *Angewandte Chemie* (30 papers)—published between 1983 and 2024.

LocalMapper (Chen et al., 2024) was applied to all reaction SMILES for atom mapping and reaction template extraction. Reagent SMILES were obtained by converting reagent names to SMILES using OPSIN (EMBL-EBI, 2026) and the PubChem API (National Center for Biotechnology Information, 2026). To quantify route complexity, we computed the longest linear sequence (LLS) for each total synthesis route using the NetworkX graph library (NetworkX developers, 2026). Specifically, for each publication, we constructed a directed graph linking reactants and products across the reported reaction sequence, and defined the LLS as the length of the longest path in this graph. The resulting dataset is referred to as TotSyn.

3. Metadata & Governance

Each reaction entry contains the following fields: atom-mapped reaction SMILES, reaction templates, reagent SMILES, journal metadata, and LLS value. Since Reaxys (Elsevier, 2026) is a proprietary database, the raw data cannot be publicly released at this time.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026).

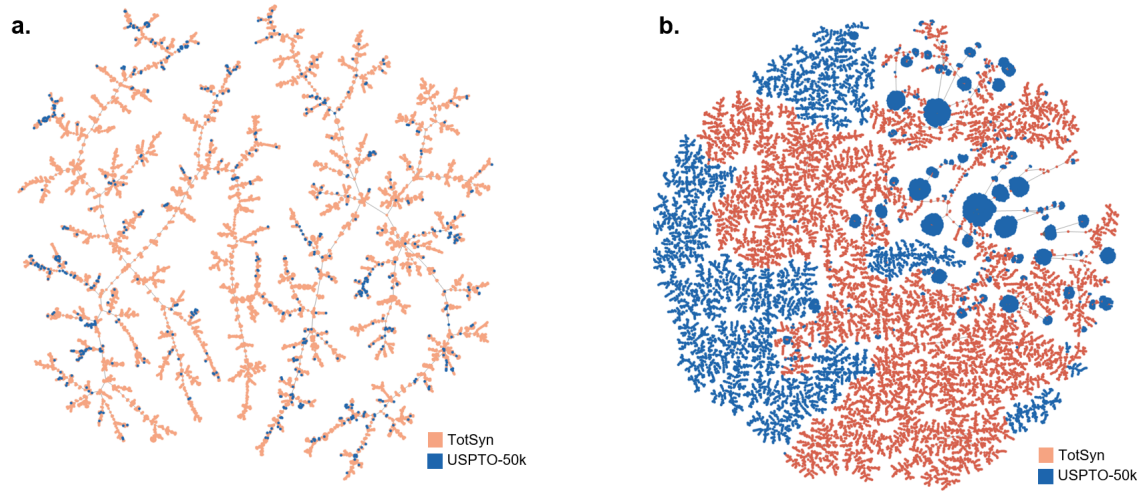

Figure 1. (a) Template space of TotSyn and USPTO-50k. (b) Reaction space of top 10k reactions of TotSyn and USPTO-50k.

4. Data Analysis

We compare TotSyn against two existing datasets: USPTO-50k and PaRoutes (Genheden & Bjerrum, 2022). PaRoutes is a route-level dataset derived from USPTO, providing multi-step synthesis routes that we utilize for LLS computation. As shown in Table 1, the resulting dataset contains 105,002 reactions and 30,703 reaction templates. Compared with existing datasets, TotSyn is larger than USPTO-50k in reaction count, though smaller than PaRoutes. Despite having fewer reactions than PaRoutes, TotSyn contains substantially more reaction templates than both, with nearly three times more than PaRoutes and nearly 13 times more than USPTO-50k, reflecting the high chemical diversity inherent to total synthesis.

To analyze template and reaction space, we adopted the reaction fingerprint embedding method proposed by Schwaller et al. (Schwaller et al., 2021). Before visualization, we removed templates that appeared only once in each dataset, as these highly rare cases can act as outliers. This filtering resulted in 9,824 templates for TotSyn and 803 templates for USPTO-50k. Figure 1(a) shows the template distributions of the two datasets, and Figure 1(b) shows 10,000 sampled reactions drawn from templates ranked in descending order of reaction frequency. Overall, TotSyn is more evenly distributed across both template space and reaction space, whereas USPTO-50k is more strongly clustered in specific regions. This highlights the broader diversity of total synthesis chemistry compared to the more concentrated reaction patterns captured in patent-derived datasets.

As shown in Appendix A, TotSyn exhibits longer LLS distributions compared to PaRoutes, where 57.4% and 24.3% of PaRoutes routes are concentrated at steps 2 and 3, respectively, indicating lower route diversity than TotSyn. To further quantify template diversity, we computed the Shannon entropy and effective number of templates at each LLS step. As shown in Appendix B, TotSyn consistently exhibits higher Shannon entropy than PaRoutes up to LLS 12, and while the effective number of templates in PaRoutes decreases from 229, TotSyn continues to increase until LLS 11, suggesting that longer synthetic routes in TotSyn are supported by a broader and more diverse set of reaction templates.

5. Acceleration potential

Total synthesis remains underrepresented in machine learning resources, as existing patent-derived datasets fail to capture its chemical diversity and strategic complexity. TotSyn addresses this gap with 105,002 reactions and 30,703 unique templates curated from peer-reviewed journals, constructed through a systematic pipeline of literature retrieval, atom mapping, and LLS computation.

Comparative analysis against USPTO-50k confirms that TotSyn exhibits substantially broader template diversity and more evenly distributed reaction space. Specifically, TotSyn enables (1) benchmarking of single-step retrosynthesis models on chemically diverse reactions, (2) training of route-level planning models that generalize beyond patent chemistry, and (3) evaluation of LLM-based synthesis agents on realistic multi-step planning tasks.

References

Chen, S., An, S., Babazade, R., and Jung, Y. Precise atom-to-atom mapping for organic reactions via human-in-the-loop machine learning. *Nature Communications*, 15(1): 2250, 2024.

Dai, H., Li, C., Coley, C., Dai, B., and Song, L. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32, 2019.

Deng, Y., Zhao, X., Sun, H., Chen, Y., Wang, X., Xue, X., Li, L., Song, J., Hsieh, C.-Y., Hou, T., et al. Rsgpt: a generative transformer model for retrosynthesis planning pre-trained on ten billion datapoints. *Nature communications*, 16(1):7012, 2025.

Elsevier. Reaxys. https://www.reaxys.com/, 2026. Chemistry database. Accessed: 2026-04-20.

EMBL-EBI. Opsin: Open parser for systematic iupac nomenclature. https://www.ebi.ac.uk/opsin/, 2026. OPSIN webserver. Accessed: 2026-04-20.

Genheden, S. and Bjerrum, E. Paroutes: towards a framework for benchmarking retrosynthesis route predictions. *Digital Discovery*, 1(4):527–539, 2022.

National Center for Biotechnology Information. Pug rest - pubchem. https://pubchem.ncbi.nlm.nih.gov/docs/pug-rest, 2026. Documentation page. Accessed: 2026-04-20.

NetworkX developers. Networkx documentation. https://networkx.org/en/, 2026. Software documentation. Accessed: 2026-04-20.

Sathyanarayana, S. V., Hiremath, S. D., Shah, R., Panda, R., Jana, R., Singh, R., Irfan, R., Murali, A., and Ramsundar, B. Deepretro: Retrosynthetic pathway discovery using iterative llm reasoning. *arXiv preprint arXiv:2507.07060*, 2025.

Schwaller, P., Probst, D., Vaucher, A. C., Nair, V. H., Kreutter, D., Laino, T., and Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature machine intelligence*, 3(2):144–152, 2021.

Tyrin, A. S., Boiko, D. A., Kolomoets, N. I., and Ananikov, V. P. Digitization of molecular complexity with machine learning. *Chemical Science*, 16(16):6895–6908, 2025.

Zheng, S., Zeng, T., Li, C., Chen, B., Coley, C. W., Yang, Y., and Wu, R. Deep learning driven biosynthetic pathways navigation for natural products with bionavi-np. *Nature Communications*, 13(1):3342, 2022.

A. LLS Distribution

Figure 2 shows the LLS distribution of TotSyn and PaRoutes. The left y-axis represents the frequency of each LLS step, while the right y-axis indicates the cumulative number of reactions.

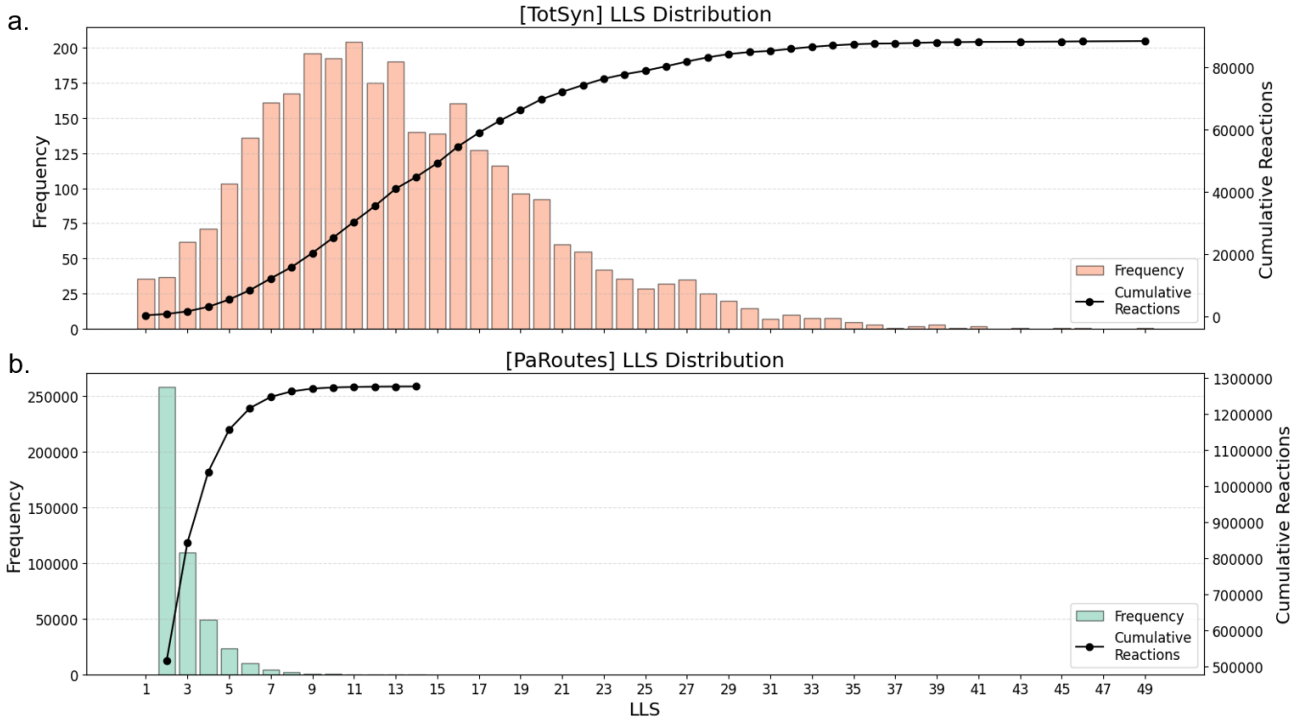

Figure 2. (a) LLS distribution of TotSyn. (b) LLS distribution of PaRoutes.

B. Template Diversity

Figure 3(a) shows the normalized Shannon entropy of reaction template distributions across LLS steps for TotSyn and PaRoutes. The normalized entropy ranges from 0 to 1, where values close to 0 indicate that a small number of templates dominate the distribution (low diversity), whereas values close to 1 indicate a more uniform distribution across observed templates (high diversity). The normalized Shannon entropy is defined as

$$H_{\text{norm}} = \frac{H}{\ln K} \in [0, 1],$$

where

$$H = -\sum_i p_i \ln(p_i),$$

p_i denotes the probability of template i, and K is the number of observed templates.

Figure 3(b) shows the effective number of templates at each LLS step, defined as

$$N_{\text{eff}} = e^H.$$

While H_{norm} measures how evenly templates are used, N_{eff} indicates how many templates are effectively in use at each LLS step. An increasing trend in N_{eff} with LLS suggests that longer synthetic routes rely on a broader set of reaction templates.

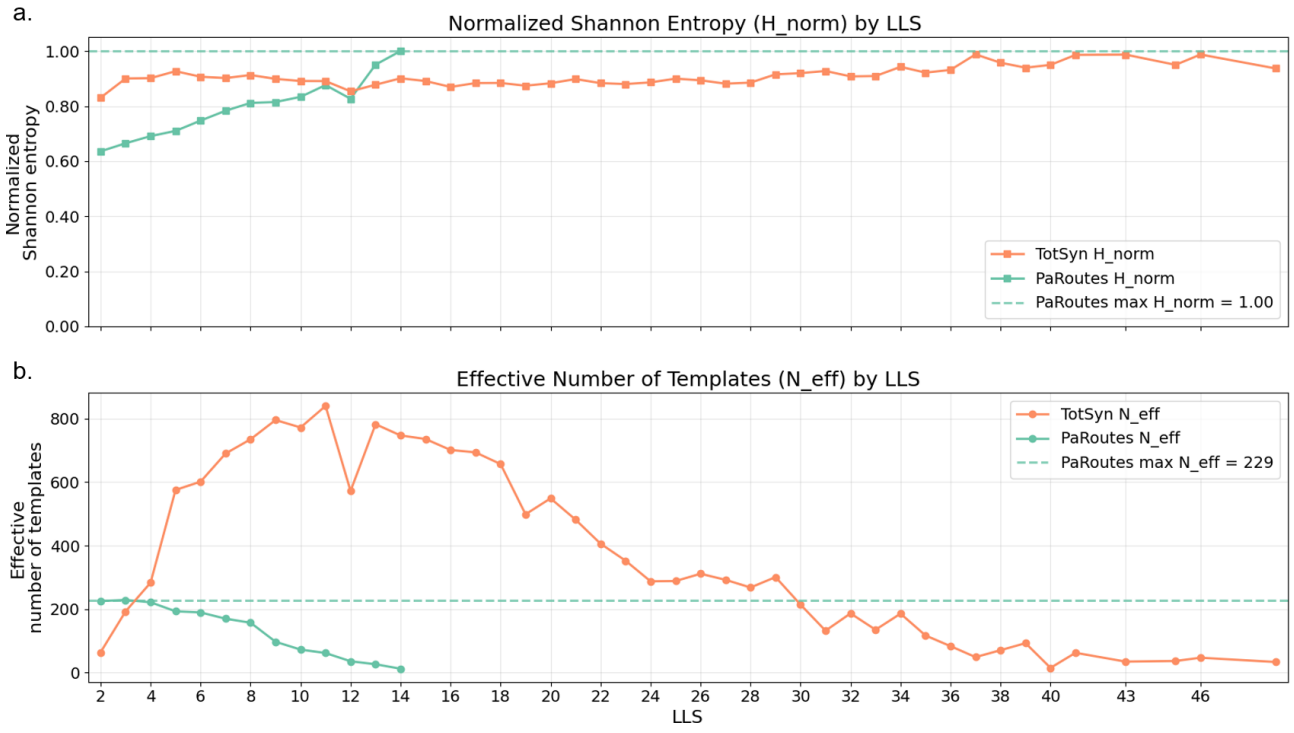

Figure 3. (a) Comparison of Shannon entropy between TotSyn and PaRoutes. (b) Compariosn of effective number of templates between TotSyn and PaRoutes.