SynthonGPT Mini v0.1.0 Technical Report

Miroslav Lžičař
DeepMedChem
miroslav.lzicar@deepmedchem.com

November 2025

Keywords: Chemical space, molecular similarity search, generative retrieval, GPT, synthons

Abstract

We introduce *SynthonGPT*, a compact synthon-conditioned transformer inspired by GPT architecture for virtual screening and molecular search in ultra-large chemical spaces. Using three $N = 100$ query sets (random samples from ChEMBL 33, Freedom Space 4.0, and ZINC) we compare SynthonGPT against F-Trees and SpaceLight. Fair, count-matched evaluations show that SynthonGPT recovers substantially more unique Bemis–Murcko scaffolds per query (up to 3.1× vs. F-Trees and 1.76× vs. SpaceLight) while maintaining lower mean ECFP4 (radius 2, 2048-bit) Tanimoto similarity to the query (i.e., higher diversity). A high-level model overview is provided without disclosing proprietary implementation details, and a live prototype is available at https://synthongpt.mireklzicar.com. Results suggest SynthonGPT complements established vendor tools by widening scaffold and structure exploration at production scale. We refer to Freedom as Chemspace Freedom Space 4.0 [1, 6].

1 Executive summary

This note summarizes the November 2025 internal readout for SynthonGPT Mini v0.1.0, detailing the system that was built, the benchmarking protocol, and deployment considerations. Ultra-large chemical space navigation is still dominated by similarity search engines such as F-Trees [7] and SpaceLight [3]. SynthonGPT complements these tools by emphasizing *diversity-oriented retrieval*: it broadens Bemis–Murcko scaffold coverage [4] while controlling for ECFP4 Tanimoto similarity and diversity [8]. All results below rely on

three 100-query samples from ChEMBL 33, Freedom Space 4.0, and ZINC with strict count matching so every method contributes the same number of hits per query. Headline result: SynthonGPT delivers up to 3.1× more unique scaffolds than F-Trees and 1.76× more than SpaceLight while slightly lowering mean ECFP4 similarity.

2 SynthonGPT Mini overview

Figure 1 shows the deployed topology; the prose below focuses on operational facts rather than academic novelty.

Inputs. Product SMILES are embedded once via a frozen internal graph encoder with one embedding per every atom of the input molecule. Gradients do not flow back to the encoder during SynthonGPT training, which keeps the upstream IP stable and limits the amount of data engineering required for the decoder work.

Decoder. A 6-layer, 8-head Transformer decoder [9] with 512 model width emits a reaction token and up to four synthon tokens. Shared embeddings and sinusoidal positional encodings are retained from earlier prototypes to minimize memory.

Outputs. Two linear heads (reaction + synthon) sit on top of the decoder. Slot-aware masking keeps synthons consistent with the predicted reaction class and optional empty slots allow shorter decompositions.

Scale and inference. The shipped configuration contains roughly 90M trainable parameters. Training runs about ten hours on a single RTX 4090 with AdamW, cosine decay, and standard cross-entropy losses. At inference time we use nucleus sampling ($p = 0.95$) with temperature 0.7 for synthons plus a width-4 beam search on the reaction token. Additional regularizers, filtering heuristics, and data curation choices remain internal by design. The model has subsecond inference speed on both GPU and CPU, requiring minimal compute costs.

3 Evaluation setup

The benchmarking harness is intentionally straightforward so results can be reproduced by other internal teams:

Baselines. Vendor tools include F-Trees 7.0.0 and SpaceLight 2.0.0 with their default CLI recipes plus explicit flags to standardize result counts: `--max-nof-results 100` and `--min-similarity-threshold 0.0`

for both tools; for F-Trees we also set `--total-diversity 1.0` (vendor-recommended default and maximum for diversity). These settings ensure exactly 100 nearest neighbors per query without early cutoffs (see Appendix B. Both baselines and SynthonGPT operate on the same Chemspace Freedom Space 4.0 enumerated chemical space.

Queries. Three query sets are sampled by taking 100 random molecules each from ChEMBL 33, Freedom, and ZINC. Some queries may be dropped if any method fails to produce hits, so the effective count per dataset can be slightly lower. In practice the intersection of queries with valid hits from all three methods is 88, 84, and 95 for ChEMBL 33, Freedom, and ZINC, respectively (see Appendix C).

Count matching. For every surviving query we take the minimum available hit count across the three methods and downsample the larger hit lists to that number. This prevents a method from "winning" by returning more molecules and is deliberately generous to the vendor baselines: queries that F-Trees or SpaceLight fail to process are removed from the evaluation rather than counted as misses, whereas SynthonGPT registers no such errors and no hallucinations (all generated molecules map back into the underlying Chemspace Freedom Space enumeration).

Data sources and leakage control. Training reactions/products are sampled from the Freedom Space 4.0 enumerated chemical space using its synthon-format schema. The Freedom query set is drawn from the same space; ChEMBL 33 and ZINC represent out-of-domain chemistry. To avoid evaluation leakage, all molecules in the three query sets are held out from training (0% overlap). Synthon format, building blocks, and reaction definitions used for training are available from Chemspace upon request.

Canonicalization and overlap. We canonicalize SMILES with RDKit (stereo-aware) before computing overlaps and diversity metrics. Deduplication is performed *within* each method's hit list; therefore overlap between F-Trees and SpaceLight is expected and reported explicitly, and method-overlap figures are computed on RDKit canonical SMILES.

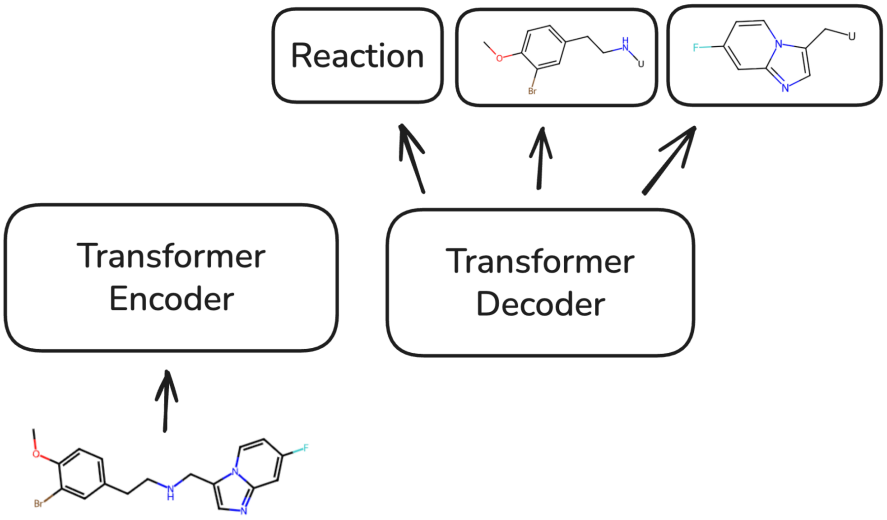

Figure 1: High-level SynthonGPT flow. A frozen product encoder provides a fixed vector to a small Transformer decoder which predicts a reaction token and up to four synthon tokens with slot-aware masking. Sensitive implementation details are intentionally omitted.

Metrics. We track (i) mean ECFP4 Tanimoto (radius 2, 2048-bit) to the query [8, 2], (ii) distinct Bemis–Murcko scaffolds [4], and (iii) hit-set overlap. These metrics align with how medicinal chemists evaluate space coverage [5].

4 Benchmark highlights

The table above drives the main narrative: SynthonGPT roughly doubles scaffold coverage on ChEMBL (2.00× vs. F-Trees; 1.74× vs. SpaceLight), triples it on Freedom (3.11×; 1.76×), and maintains a solid lead on ZINC (1.90×; 1.36×). Crucially, mean ECFP4 similarity drops instead of rising, e.g. Freedom: 0.388 ± 0.162 for SynthonGPT vs. 0.500 ± 0.153 (F-Trees) and 0.639 ± 0.127 (SpaceLight); ChEMBL: 0.207 ± 0.098 vs. 0.269 ± 0.144 and 0.373 ± 0.165; ZINC: 0.293 ± 0.119 vs. 0.307 ± 0.122 and 0.480 ± 0.116 (see Appendix D for the full table). The extra scaffolds are therefore not gained by drifting closer to the queries.

Figures 2 and 3 provide the supporting detail expected in a technical

Dataset	Method	Num queries	N	Murcko scaffolds ↑	Mean Tanimoto ↓
Chembl 33	FTrees	88	8633	1817	0.269
	SpaceLight	88	8633	2097	0.373
	SynthonGPT	88	8633	**3639**	**0.207**
Freedom 140B	FTrees	84	8264	1111	0.500
	SpaceLight	84	8264	1966	0.639
	SynthonGPT	84	8264	**3456**	**0.388**
Zinc 22	FTrees	95	9363	2495	0.307
	SpaceLight	95	9363	3490	0.480
	SynthonGPT	95	9363	**4736**	**0.293**

Table 1: Count-matched diversity summary across three query sets. N is the total pooled number of matched hits considered after per-query downsampling. Bold highlights the best value within each dataset block. Trends favor SynthonGPT on scaffold coverage and diversity while maintaining low similarity.

report: they show the pooled similarity distributions (after count-matching on the 88/84/95-query intersections) and per-query scaffold counts that guide downstream workflow planning.

Limitations. The present snapshot covers a single vendor space and three random query samples. More datasets—especially focused on project-specific chemical liabilities—are needed before a broad external release. Architectural, data, and filtering specifics remain intentionally redacted to protect proprietary work.

An orthogonality check across the pooled hit sets highlights how SynthonGPT complements the vendor searches. F-Trees contributes 24,264 unique SMILES, SpaceLight returns 27,244, and their shared overlap is 2,335. SynthonGPT adds 29,320 hits that are non-overlapping with both vendor lists under RDKit canonical SMILES (stereo-aware), confirming that the decoder surfaces chemotypes that current tools cannot reach under identical count-matched constraints (Figure 4).

5 Conclusion

SynthonGPT Mini v0.1.0 behaves like a pragmatic front-end for diversity-oriented retrieval. Under matched evaluations it widens scaffold coverage while lowering ECFP4 similarity relative to incumbent tools across three query sets, reinforcing its value as part of a multi-method discovery flow.

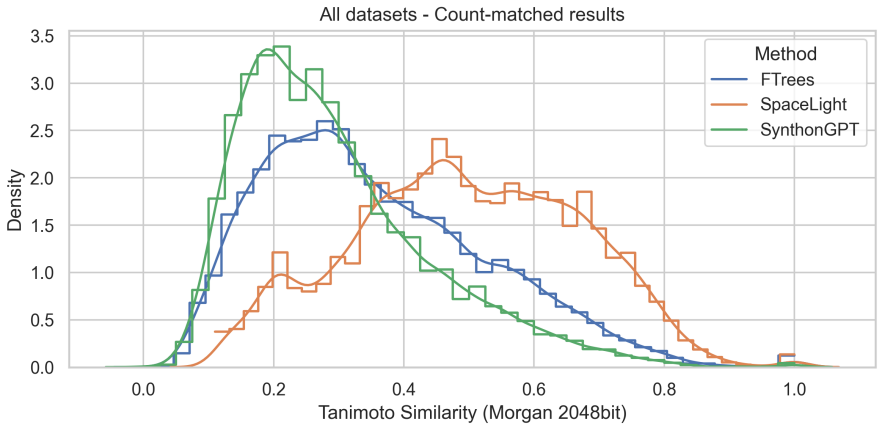

Figure 2: Count-matched Tanimoto distributions (pooled). F-TREES concentrates near lower diversity; SPACELIGHT shifts toward higher similarity; SynthonGPT maintains a broader low-similarity tail, indicating more diverse hits.

References

[1] Freedom space 4.0. `https://chem-space.com/freedom-space`, 2025. Chemspace; accessed November 2025.

[2] Bajusz, D., Rácz, A., and Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20, 2015. doi: 10.1186/s13321-015-0069-3.

[3] Bellmann, L., Penner, P., and Rarey, M. Topological similarity search in large combinatorial fragment spaces. *Journal of Chemical Information and Modeling*, 61(1):238–251, 2021. doi: 10.1021/acs.jcim.0c01068.

[4] Bemis, G. W. and Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

[5] Hoffmann, T. and Gastreich, M. The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today*, 24(5):1148–1156, 2019. doi: 10.1016/j.drudis.2019.02.013.

[6] Kapeliukha, A., Hlotov, S., Protopopov, M., Dzyuba, I., Vasylchuk, M.,

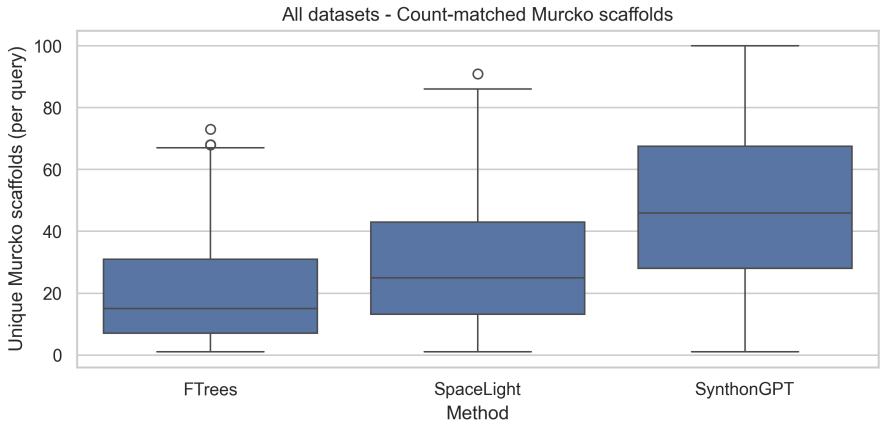

Figure 3: Distinct Bemis–Murcko scaffolds per query (count-matched). SynthonGPT attains the highest median (≈ 45/query) and the widest upper quartile; SpaceLight is intermediate; F-Trees trails.

Panov, D. M., Tarkhanova, O. O., and Moroz, Y. S. Freedom space 3.0: Ml-assisted selection of synthetically accessible small molecules. *Journal of Chemical Information and Modeling*, 65(19):10338–10347, 2025. doi: 10.1021/acs.jcim.5c01912.

[7] Rarey, M. and Dixon, J. S. Feature trees: a new molecular similarity measure based on tree matching. *Journal of Computer-Aided Molecular Design*, 12(5):471–490, 1998. doi: 10.1023/A:1008068904628.

[8] Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A Supplementary notes

Baseline command-line details are provided in the supplement. Figures and table use count-matched subsets to normalize per-query hit counts before aggregation.

B Baseline CLI and reproducibility

CLI flags. Runs used vendor defaults plus: `--max-nof-results 100` and `--min-similarity-threshold 0.0` for both F-Trees and SpaceLight; for F-Trees we set `--total-diversity 1.0` (vendor-recommended and maximum diversity).

Fingerprints. ECFP4 computed with RDKit (Morgan, radius 2, 2048 bits; chirality enabled). SMILES canonicalization uses RDKit canonical SMILES (stereo-aware).

Commands. For reference:

```
# FTrees (schematic)
FTrees -i "c1ccccc1" \
       -s FreedomSpace_140bn_2025-07.space \
       --max-nof-results 100 \
       --min-similarity-threshold 0.0 \
       --total-diversity 1.0
# SpaceLight (schematic)
SpaceLight -i "c1ccccc1" \
       -s FreedomSpace_140bn_2025-07.space \
       --max-nof-results 100 \
       --min-similarity-threshold 0.0
```

C Error analysis and query filtering

Error logs from the FreedomSpace_140bn_2025-07 runs explain why the effective query counts in the diversity plots are slightly below 100 and clarify the robustness of each method. Across the three datasets, F-Trees accumulates 32 failed queries (24 timeouts, 6 macrocycle rejections, and 2 invalid SMILES) and SpaceLight reports 2 invalid-SMILES failures on the same organotin-containing molecules. In contrast, SynthonGPT completes all queries without errors. The shared intersection of queries with valid hits from all three methods is therefore 88/84/95 for ChEMBL 33/Freedom/ZINC. Restricting the benchmark to this intersection makes the count-matched comparison both fair and generous to the baselines: queries that the vendor engines cannot process are dropped entirely rather than counted as failures, even though SynthonGPT succeeds on them. We additionally verified that 100% of SynthonGPT hits correspond to enumerated molecules in the Chemspace Freedom Space, so the model exhibits zero hallucinations in this benchmark.

Dataset	Method	Timeouts	Macrocycles	Invalid SMILES	Total
ChEMBL 33	F-Trees	6	5	0	11
	SpaceLight	0	0	0	0
	SynthonGPT	0	0	0	0
Freedom	F-Trees	13	1	2	16
	SpaceLight	0	0	2	2
	SynthonGPT	0	0	0	0
ZINC	F-Trees	5	0	0	5
	SpaceLight	0	0	0	0
	SynthonGPT	0	0	0	0

Table 2: Query-level failures by dataset and method derived from the vendor error CSVs for the FreedomSpace_140bn_2025-07 runs. SynthonGPT completes all 300 original queries with no logged errors.

D Full metrics table

Dataset	Method	Num queries	N	Murcko ↑	Mean Tan. ↓	SD
Chembl 33	FTrees	88	8633	1817	0.269	0.144
Chembl 33	SpaceLight	88	8633	2097	0.373	0.165
Chembl 33	**SynthonGPT**	88	8633	**3639**	**0.207**	**0.098**
Freedom 4.0	FTrees	84	8264	1111	0.500	0.153
Freedom 4.0	SpaceLight	84	8264	1966	0.639	0.127
Freedom 4.0	**SynthonGPT**	84	8264	**3456**	**0.388**	**0.162**
Zinc 22	FTrees	95	9363	2495	0.307	0.122
Zinc 22	SpaceLight	95	9363	3490	0.480	0.116
Zinc 22	**SynthonGPT**	95	9363	**4736**	**0.293**	**0.119**

Table 3: Full count-matched metrics including standard deviation (SD) of Tanimoto similarity. N is the pooled number of matched hits after per-query downsampling.

E Tool versions and build dates

Table 4 summarizes the exact vendor tool builds referenced in this report together with the SynthonGPT release and chemical space snapshot evaluated in the benchmark.

Tool / Space	Version	Build / Snapshot
SynthonGPT Mini	0.1.0	2025-11-16
BioSolveIT FTrees	7.0.0	2025-07-28 15:15
BioSolveIT SpaceLight	2.0.0	2025-07-28 15:18
Chemspace Freedom Space	4.0	140B snapshot (2025-07)

Table 4: Software builds and chemical space snapshot used in the study.

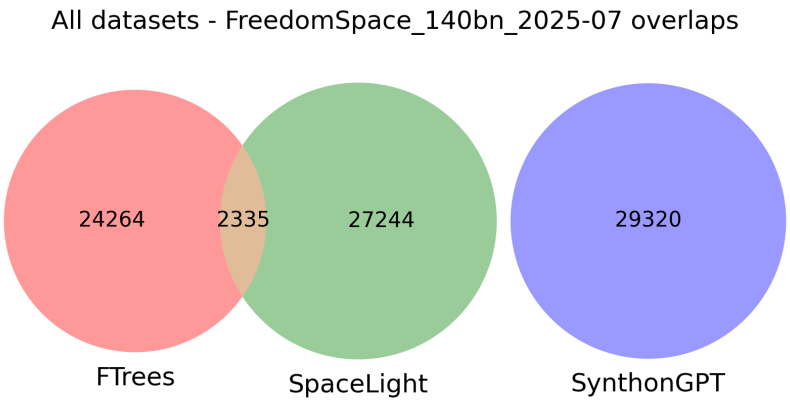

Figure 4: Unique SMILES overlap across methods (count-matched evaluation) computed on RDKit canonical SMILES (stereo-aware). F-Trees: 24,264, SpaceLight: 27,244, shared F-Trees/SpaceLight: 2,335, SynthonGPT: 29,320 (non-overlapping with both vendor lists).