Factorized Search and Cartography of Synthon-Based Chemical Spaces

Miroslav Lžičař [1]

Abstract

Make-on-demand molecular design spaces are increasingly defined by reactions and synthons rather than by enumerated molecules, creating a grounding problem for synthesis-aware discovery workflows that must search, audit, and visualize spaces too large to materialize. We present a simple factorized framework with two complementary operations over a shared synthon-product surrogate. SYNTHONOR approximates each candidate product fingerprint as the bitwise OR of its participating synthon fingerprints and retrieves query-nearest tuples by Tanimoto ranking. SYNTHONMAP uses the same surrogate for route-aware sampling, importance weighting, nonlinear embedding, and product-mass cartography. On enumerable RDKIT synthon-space slices, raw-connector ECFP4 OR similarities preserve assembled-product similarity rankings with Spearman $\rho = 0.928$ and recover 63.4% mean exact top-100 overlap over one query per reaction. At Freedom Space 4.0 scale, SYNTHONOR completes all 100 benchmark queries and gives direct ECFP4 hit-similarity comparisons with FTrees and SpaceLight, while SYNTHONMAP visualizes a million-product sample with importance-weighted density. The result is a compact, transparent, and practical search-and-audit layer for factorized chemical spaces, designed to complement synthesis-aware generators and library-inspection workflows.

1. Introduction

Synthesis-aware molecular-design systems increasingly operate at the boundary between biology, molecular design, and make-on-demand chemistry. Synthesis-aware generators such as PrexSyn, SynFormer, SynFlowNet, and SynthemMol move molecular generation toward reaction- and building-block-conditioned design spaces (Luo & Coley, 2025; Gao et al., 2025; Cretu et al., 2025; Swanson et al., 2024). A model may propose a motif, an automated workflow may need to expand a hit, and a chemist may need to inspect whether a vendor-defined space contains useful nearby chemistry. In all three cases the actionable domain is often not a fixed molecular table but a factorized reaction space: a set of routes, each with slots populated by compatible synthons. Modern tangible spaces can contain billions to trillions of products (Walters, 2019; Warr et al., 2022; Tingle et al., 2023; Patel et al., 2020; Kapeliukha et al., 2025; Liu et al., 2025), and Chemspace's current Freedom Space release is marketed at hundreds of billions of molecules (Chemspace, 2026), so exhaustive enumeration is frequently the wrong primitive.

Two practical questions recur. First, given a query molecule, which products in the factorized space are nearest neighbors? Second, before asking a query, what does the product universe look like, where is its mass, and which routes dominate or diversify the space? Existing fragment-space search methods answer related retrieval questions with feature trees, topological fragment spaces, maximum common substructure logic, weighted fingerprints, shape-aware synthon search, or substructure-specific pruning (Rarey & Stahl, 2001; Bellmann et al., 2021; Schmidt et al., 2022; Cheng & Beroza, 2024; Liphardt & Sander, 2023; Lübbers et al., 2026). RDKit now also exposes experimental synthon-space search for exact, substructure, and fingerprint-similarity use cases, and Chemspace advertises RDKit-backed synthon search for Freedom Space files (RDKit Developers, 2026; Chemspace, 2024). Structure-based synthon workflows such as V-SYNTHES demonstrate a target-specific route: hierarchical docking over synthon-based virtual libraries (Sadybekov et al., 2022).

Chemical-space visualization has a complementary history. Self-organizing maps are a classic cheminformatics visualization model (Kohonen, 1990; Digles & Ecker, 2011), and tools such as Scaffold Hunter, CheS-Mapper, ChemPlot, TMAP, ChemSpace Atlas, and visual-navigation surveys provide strong options once molecules or representative samples are available (Medina-Franco et al., 2008; Wetzel et al., 2009; Gütlein et al., 2012; Cihan Sorkun et al., 2022; Probst & Reymond, 2020; Zabolotna et al., 2022a; Sosnin, 2025). Closest to our cartography setting, synthon

[1]Deep MedChem. Correspondence to: Miroslav Lžičař <lzicar2000@gmail.com>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

and DNA-encoded-library spaces have been analyzed with topographic-map chemography and neural surrogates, including a universal synthon-space map, chemical-library-space and Meta-GTM analyses for DELs, and CoLiNN, which predicts combinatorial-library projections without full enumeration (Zabolotna et al., 2022b; Pikalyova et al., 2022; 2023a;b; 2025). Our niche is a minimal, transparent OR-surrogate retrieval and cartography layer: local enough to audit, fast enough to run on factorized spaces, and simple enough to serve as infrastructure for generators or human library inspection.

The core idea is to approximate a product fingerprint by the binary OR of the fingerprints of the synthons that would form it. This is an inspectable surrogate for inherited local features, not a replacement for product assembly when exact chemistry is required. We use it in two modes. SYNTHONOR is the retrieval engine: query fingerprint to ranked synthon tuples. SYNTHONMAP is the atlas: synthon space to sampled, weighted, two-dimensional density views. The contribution of this joint paper is the common factorized representation, the two algorithms built on it, and evidence that the same representation supports exact-slice validation, vendor-scale retrieval snapshots, and sampled chemical-space cartography. The intended role is complementary to synthesis-aware generators: a deterministic grounding and audit layer that such systems can call, benchmark against, or use for distributional inspection.

2. Factorized Product Spaces

Definition 1 (Synthon product space)**.** Let $\mathcal{R}$ be a set of reaction routes. Each route $r \in \mathcal{R}$ has arity k_r and nonempty slot sets $S_{r,1}, \ldots, S_{r,k_r}$. The non-enumerated product universe is the disjoint union

$$\Omega = \bigsqcup_{r \in \mathcal{R}} \prod_{p=1}^{k_r} S_{r,p}. \tag{1}$$

An element $c = (r, s_1, \ldots, s_{k_r}) \in \Omega$ is a synthetically compatible synthon tuple, not necessarily an assembled molecule stored on disk.

Let $f(x) \in \{0,1\}^d$ be a binary molecular fingerprint, such as an ECFP fingerprint (Rogers & Hahn, 2010), and let $Q = f(q)$ be the query fingerprint. For a candidate tuple $c = (r, s_1, \ldots, s_{k_r})$, define the OR surrogate

$$\hat{f}(c) = \bigvee_{p=1}^{k_r} f(s_p), \tag{2}$$

where $\vee$ is bitwise OR. The same $\hat{f}$ supports both retrieval and mapping. Retrieval optimizes

$$\mathrm{topK}_{c \in \Omega}\ \mathrm{T}(Q, \hat{f}(c)), \quad \mathrm{T}(a,b) = \frac{|a \cap b|}{|a \cup b|}, \tag{3}$$

while cartography studies the pushforward measure

$$c \sim \mu, \quad \hat{f}(c) \mapsto g(\hat{f}(c)) \in \mathbb{R}^2 \tag{4}$$

for an embedding function g, followed by weighted rasterization in the learned coordinates.

Proposition 1 (Surrogate perturbation)**.** *For a fixed query bitset Q, let U be the exact product fingerprint and $\hat{U}$ its OR surrogate. If their symmetric difference has size $\delta = |U \triangle \hat{U}|$, then*

$$|\mathrm{T}(Q,U) - \mathrm{T}(Q,\hat{U})| \le \frac{2\delta}{\min\{|Q \cup U|,\ |Q \cup \hat{U}|\}}. \tag{5}$$

The proof is in Appendix A.2. The bound is intentionally modest: when the OR surrogate differs from the exact product fingerprint by few query-relevant bits relative to the query-union size, Tanimoto scores cannot move arbitrarily. In practice the method still needs empirical validation because small score perturbations can alter ranks near the decision boundary, and reaction junctions may create or delete important bits. This motivates exact-product overlap benchmarks for search and exact-slice neighborhood or density checks for maps.

3. Algorithms

3.1. Retrieval: SYNTHONOR

SYNTHONOR avoids full enumeration by combining global synthon prescreening with bounded tuple rescoring. It first precomputes packed RDKIT-compatible fingerprints and bit counts for all synthons. For query bitset Q, each synthon s receives a prescreen score $a(s) = \mathrm{T}(Q, f(s))$. For route r and slot p, let $B_{r,p}(Q)$ be the top B synthons in $S_{r,p}$ by $a(s)$. The route pre-score is

$$A_r(Q) = \sum_{p=1}^{k_r} \max_{s \in B_{r,p}(Q)} a(s). \tag{6}$$

The implementation keeps the top C routes by A_r (or all routes when $C = \infty$). For each retained route, it enumerates the Cartesian product of the per-slot banks; if the bank product exceeds the tuple budget E, each slot is truncated to a balanced prefix so that the retained product has at most E tuples. Each tuple is then rescored by $\mathrm{T}(Q, \bigvee_p f(s_p))$, sorted globally, and filtered to the requested top-K hits. Product assembly is optional downstream, but it is absent from the scoring loop.

Let $N = \sum_{r,p} |S_{r,p}|$, $w = \lceil d/8 \rceil$ packed bytes, $k_{\max} = \max_r k_r$, and M be the number of materialized tuples after route and budget caps, with $M \le \sum_{r \in \mathrm{kept}} \min(E, \prod_p |B_{r,p}|)$. The current exact-sorting implementation has per-query cost $O(Nw +$

Table 1. Surrogate fidelity and 10-query fingerprint-family retrieval ablation. Fidelity uses the raw-connector RDKIT enumerable slice with 20,000 random product pairs and 29,200 products; retrieval uses the deterministic 990,829-product slice with 10 queries. Retrieval overlap is mean exact top-100 overlap; the 42-query ECFP4 all-reaction check is reported in Table 5.

Fingerprint	Pair ρ	Pair MAE	Prod. median T	Exact-bit cov.	OR prec.	Fast	Fast s	Accurate	Accurate s
ECFP4	0.928	0.030	0.584	0.774	0.700	56.2	0.963	63.9	7.337
ECFP6	0.876	0.031	0.470	0.644	0.632	49.7	0.925	56.6	7.376
Topological torsion	0.884	0.020	0.642	0.733	0.832	30.2	0.874	33.4	7.455
RDKIT	0.866	0.087	0.490	0.560	0.783	20.9	1.110	25.4	7.520
Atom pair	0.798	0.038	0.311	0.332	0.838	9.8	1.202	13.8	7.572
PatternFP	0.779	0.136	0.524	0.577	0.868	2.2	1.215	2.2	7.509
Morgan counts (max)	0.913	0.017	0.570	0.705	0.743	–	–	–	–
Morgan counts (sum)	0.982	0.030	0.647	0.824	0.747	–	–	–	–

Table 2. SYNTHONOR search presets in the released implementation. $C = \infty$ denotes all prescreened routes; E is the per-route tuple budget before final global sorting.

Preset	C	B	E	Default K
Fast	8	64	50k	100
Accurate	∞	192	250k	100
Very accurate	∞	256	500k	100

$\sum_{r,p} |S_{r,p}| \log |S_{r,p}| + Mk_{\max}w + M \log M)$ and memory $O(Nw + M)$ for packed binary fingerprints. Its moving parts are intentionally visible: fingerprint family, candidate-bank size, route cap, tuple budget, and union mode. The baseline is training-free and does not require connector heuristics, fragment graph matching, or reaction execution during ranking.

3.2. Cartography: SYNTHONMAP

SYNTHONMAP changes the scientific question from nearest neighbors to product-space measure. For each route r, let

$$C_r = \prod_{p=1}^{k_r} |S_{r,p}|, \quad C_{\text{tot}} = \sum_{r \in \mathcal{R}} C_r. \tag{7}$$

Sampling first chooses a route and then one synthon uniformly from each slot. Product-uniform sampling uses $p_r = C_r/C_{\text{tot}}$ and gives each product equal probability. Reaction-balanced sampling uses $p_r = 1/|\mathcal{R}|$ to improve route coverage. Square-root-balanced sampling uses $p_r \propto \sqrt{C_r}$ as a compromise.

When sampling is not product-uniform, a sampled product from route r receives an importance weight proportional to

$$w(c) \propto \frac{1/C_{\text{tot}}}{p_r/C_r} = \frac{C_r}{p_r}, \tag{8}$$

up to a constant irrelevant for normalized density plots. The sampled OR fingerprints are embedded with UMAP, DENSMAP, t-SNE, or PCA (van der Maaten & Hinton, 2008; McInnes et al., 2018; Narayan et al., 2021); two-dimensional histograms can then show either unweighted sample coverage or importance-weighted product mass. This separation is important because nonlinear embedding density is a property of the learned view, not a guaranteed high-dimensional chemical density.

4. Experiments

4.1. Surrogate fidelity on assembled products

The OR surrogate is deliberately lossy, so we first ask whether it preserves similarity structure on enumerable products. We use an RDKIT synthon-space analysis with 312 synthons and 29,200 enumerated products. For 20,000 random product pairs, we compare exact assembled-product fingerprint similarity with the similarity induced by synthon-composed fingerprints. Table 1 summarizes the raw-connector setting, which retains attachment context in the synthon fingerprints, together with the downstream exact-product retrieval benchmark. ECFP4 gives strong pairwise rank agreement (Spearman $\rho = 0.928$, mean absolute error 0.030), while count Morgan fingerprints with sum composition reach $\rho = 0.982$. Product-vs-OR fingerprint agreement is lower than the pairwise-rank agreement, as expected: ECFP4 has median $\mathrm{T}(f_{\text{exact}}, \hat{f}) = 0.584$, mean exact-bit coverage 0.774, and mean OR precision 0.700. This does not make OR chemically exact, but it explains why a transparent tuple-level surrogate can recover a substantial fraction of exact nearest neighbors.

4.2. Exact-product retrieval

We make the broader all-reaction run the primary exact-retrieval check. The deterministic slice contains 6,273 synthons, 42 reactions, and 990,829 assembled products. We sample one full product from each reaction as a query, assemble and score every product in the slice as the reference, and evaluate overlap with the exact top-100 products. Table 5 shows that the connector-retaining ECFP4

policy recovers 63.4% mean exact top-100 overlap over 42 queries, with a median of 67.0 and a standard deviation of 17.1. A faster global-synthon prescreen lowers mean overlap to 59.2% while reducing mean wall time from 1.13s/query to 0.80s/query. These overlaps characterize candidate-generation quality rather than exact chemical-neighbor recovery, because the returned order is still the OR-surrogate ranking. The 10-query fingerprint-family ablation in Table 1 is therefore not the headline retrieval evidence; it shows that ECFP4 is the strongest tested binary fingerprint family under the earlier fast/accurate presets.

4.3. Freedom Space 4.0 scale-up

We next instantiate the same search machinery on the Freedom Space 4.0 benchmark used in infiniSee-style comparisons. This experiment tests operational scale and returned-hit quality when exact top-k recall cannot be measured by full enumeration of the 140B-product space. The query set is the 100-SMILES `freedom_142B_100` panel, each completed query contributes at most 100 returned products, and all hit similarities are recomputed ECFP4 Tanimoto scores between the query and returned product SMILES. Failed FTrees and SpaceLight queries are the timeout/error rows present in the merged artifacts; they are excluded from completed-query medians, included as zero top-1 similarity in the 100-query penalized column, and removed from the 84-query common-success subset. The archived FTrees and SpaceLight outputs did not retain wall times or software-version metadata, so timing is reported only for the SYNTHONOR runs.

Table 3 and Figure 1 show the resulting snapshot. SYNTHONOR fast and SYNTHONOR accurate return 100 hits for all 100 queries, while the FTrees and SpaceLight merged outputs contain 84 and 98 completed queries respectively. SYNTHONOR accurate has higher median hit Tanimoto than SYNTHONOR fast in this archive (0.684 versus 0.662), with median query time increasing from 1.44s to 17.68s. The result demonstrates that the factorized OR index can run on a vendor-scale synthon space and can be inspected against established search modes without materializing products.

4.4. Product-space maps

Figure 2 shows a SYNTHONMAP run on the same Freedom Space 4.0 source used for the retrieval snapshot, whose route-slot products imply 141.97B approximate products in the loaded synthon table. This is the full Freedom Space 4.0 140B synthon space used for the benchmark snapshot. We draw 1,000,000 products with square-root-balanced route sampling and seed 7, compute 2048-bit ECFP4 OR fingerprints, and place them in a Jaccard UMAP coordinate system fit on a 100,000-product anchor sample using 50 neighbors and minimum distance 0.02. We then render both

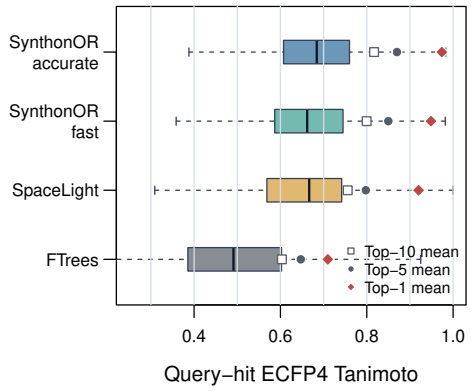

Figure 1. Returned-hit ECFP4 similarities on Freedom Space 4.0. Thin boxes summarize all returned top-100 hits; markers show mean similarity over the top 10, top 5, and top 1 hits per method.

unweighted and weighted density. The unweighted panel asks whether the exploratory sample covers the space; the weighted panel asks where product-uniform mass lies after correcting for the proposal distribution. Route coverage and concentration diagnostics are summarized in Table 6.

Table 6 adds small quantitative diagnostics for the same map machinery. On the exact 990,829-product slice, a 60,000-product square-root-balanced map covers all 42 reactions and has 15-neighbor trustworthiness 0.896 on a fixed 2,000-point Jaccard subsample. On the Freedom-scale projection, all 42 reactions are still covered, but the product-mass weights are concentrated: the largest route accounts for 77.7% of weighted mass in the sampled view. These numbers support using the figure as a route/mass audit, not as a calibrated high-dimensional density estimate.

5. Discussion

Role relative to generative models. Recent synthesis-aware generators produce molecules or routes inside learned policies. Our contribution is the complementary infrastructure layer: once a model or workflow proposes a molecule, motif, property objective, or region of interest, SYNTHONOR grounds that intent in a concrete make-on-demand universe and SYNTHONMAP exposes route coverage, product mass, and local neighborhoods. These tools therefore pair naturally with PrexSyn, SynFormer, SynFlowNet, CSearch, and related virtual-synthesis optimizers (Luo & Coley, 2025; Gao et al., 2025; Cretu et al., 2025; Kim et al., 2024): the generator proposes, while the factorized map and search layer retrieves, audits, and explains the reachable chemistry.

Biological-discovery use case. The immediate biological value is a practical grounding step for synthesis-aware de-

Table 3. Freedom Space 4.0 search snapshot on 100 Freedom-space queries. Hit similarities are recomputed ECFP4 Tanimoto values. "Common" is mean top-1 similarity on the 84 queries completed by every method; "100q" counts failed queries as zero top-1 similarity. Median times are available for the SYNTHONOR runs.

Method	Done	Err.	Median hit	Top-1 done	Top-1 common	Top-1 100q	Median s
FTrees	84	16	0.491	0.710	0.710	0.596	–
SpaceLight	98	2	0.667	0.920	0.919	0.902	–
SYNTHONOR fast	100	0	0.662	0.949	0.951	0.949	1.44
SYNTHONOR accurate	100	0	0.684	0.974	0.973	0.974	17.68

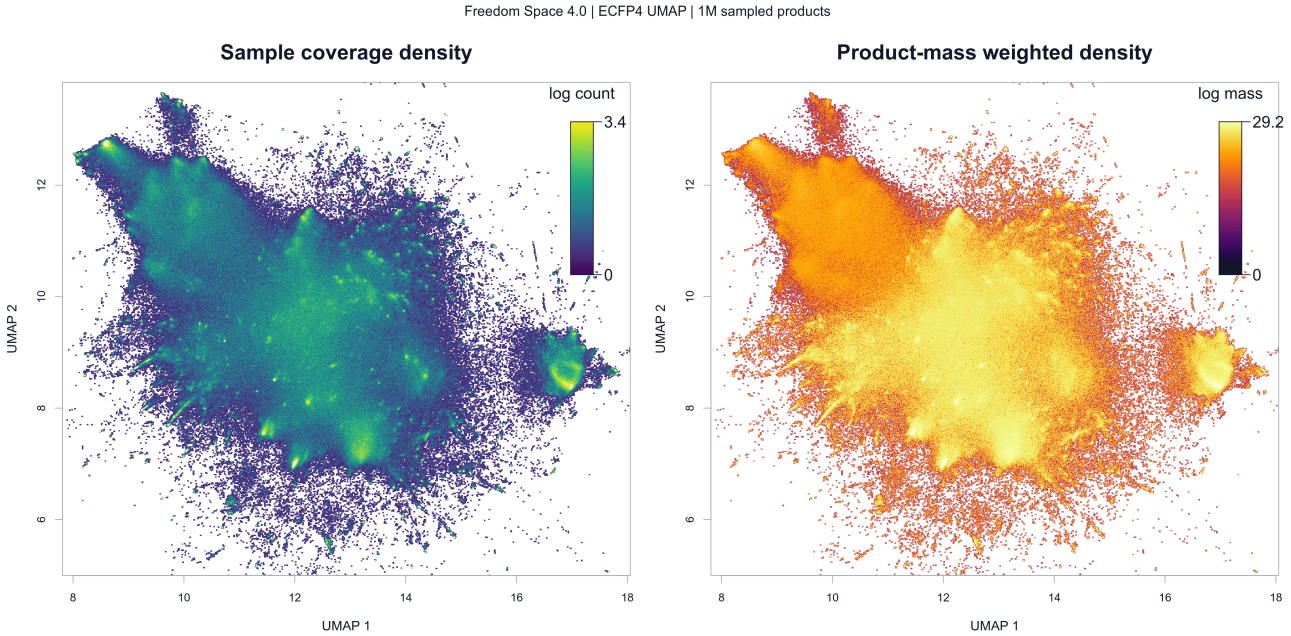

Figure 2. SYNTHONMAP map for a 1,000,000-product sample from the Freedom Space 4.0 factorized space. Left: sample coverage density. Right: product-mass weighted density under the same coordinates.

sign. A generator that proposes a kinase-like hinge-binding motif, scaffold hop, or property-biased molecule can use SYNTHONOR to retrieve purchasable analogs and SYNTHONMAP to inspect route and product-mass coverage before docking, target-specific scoring, or assay selection. This separates the deterministic library-infrastructure task from target-specific biology while making downstream validation choices easier to audit.

Limitations of binary OR. Binary OR deliberately trades chemical exactness for speed and inspectability. It inherits synthon bits and therefore does not represent reaction-junction features, disappeared synthon features, or multiplicities of repeated environments. The count-fingerprint rows in Table 1 show that count composition is promising, but an optimized count-index would be a different retrieval system from the packed-bitset implementation reported here. Candidate prescreening can also favor high-bit-density synthons, although final Tanimoto rescoring penalizes extra off-query bits. These limits suggest direct extensions: capped-count composition, bit-density-calibrated prescreening, Tversky containment scores, reaction-junction correction bits, and optional exact assembly/reranking of a bounded hit set.

Validation scope and extensions. The exact-slice retrieval benchmark and OR-vs-exact fidelity analysis quantify where the surrogate is useful, and the Freedom Space 4.0 snapshot shows that the same machinery scales to a vendor synthon space. The next validation tier is straightforward: exact assembly and reranking of bounded hit sets on enumerable slices, repeated-seed map stability, density-image correlation against exact slices, and biological recovery tests using ChEMBL actives, docking scores, or generator outputs. For maps, weighted density remains density in the learned two-dimensional display rather than a conserved high-dimensional chemical density. Clustering or coreset methods such as large-scale binary fingerprint clustering (López Pérez et al., 2025) are natural next steps for less

Table 4. Positioning relative to neighboring method classes.

Class	Examples	Main role
Generators	PrexSyn, SynFormer, SynFlowNet	sample molecules/routes
Vendor search	FTrees, SpaceLight, SpaceMACS	nearest products/substructures
Visualization	TMAP, ChemSpace Atlas, CoLiNN	map molecules/libraries
This work	SYNTHONOR + SYNTHONMAP	search and audit factorized spaces

redundant map samples.

6. Conclusion

We unified two complementary operations over synthon-based chemical spaces. SYNTHONOR uses fingerprint-OR surrogates for transparent query-conditioned retrieval, while SYNTHONMAP uses the same surrogate for route-aware sampling, weighted embedding, and density visualization. The shared factorized view turns non-enumerated reaction spaces into searchable and inspectable objects with simple ingredients: cached synthon fingerprints, bitwise OR composition, route-slot sampling, and popcount-based scoring. This makes the system a practical grounding layer for synthesis-aware molecular generators, hit-expansion workflows, and human chemical-space navigation.

Impact Statement

This work aims to improve transparent navigation of synthesizable chemical spaces for biological discovery. Potential benefits include more reproducible hit expansion, clearer auditing of make-on-demand libraries, and better baselines for molecular-design workflows. Potential risks are those common to molecular design tools, including acceleration of undesirable chemical optimization; we do not introduce biological assays, synthesis protocols, or target-specific optimization machinery in this paper.

References

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.0c00850.

Chemspace. Introducing synthon-based search for freedom space 3.0, 2024. URL `https://chem-space.com/news/synthon-search-freedom-space`. Accessed 2026-05-05.

Chemspace. Freedom Space 5.0: Ultra-large synthesizable chemical space for hit discovery, 2026. URL `https://chem-space.com/freedom-space`. Accessed 2026-05-04.

Cheng, C. and Beroza, P. Shape-aware synthon search (SASS) for virtual screening of synthon-based chemical spaces. *Journal of Chemical Information and Modeling*, 64(4):1251–1260, 2024. doi: 10.1021/acs.jcim.3c01865.

Cihan Sorkun, M., Mullaj, D., Koelman, J. M. V. A., and Er, S. ChemPlot, a python library for chemical space visualization. *Chemistry–Methods*, 2(7):e202200005, 2022. doi: 10.1002/cmtd.202200005.

Cretu, M., Harris, C., Igashov, I., Schneuing, A., Segler, M., Correia, B., Roy, J., Bengio, E., and Liò, P. SynFlowNet: Design of diverse and novel molecules with synthesis constraints. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2405.01155.

Digles, D. and Ecker, G. F. Self-organizing maps for in silico screening and data visualization. *Molecular Informatics*, 30(10):838–846, 2011. doi: 10.1002/minf.201100082.

Gao, W., Luo, S., and Coley, C. W. Generative AI for navigating synthesizable chemical space. *Proceedings of the National Academy of Sciences*, 122(41):e2415665122, 2025. doi: 10.1073/pnas.2415665122.

Gütlein, M., Karwath, A., and Kramer, S. CheS-Mapper: Chemical space mapping and visualization in 3d. *Journal of Cheminformatics*, 4:7, 2012. doi: 10.1186/1758-2946-4-7.

Kapeliukha, A., Hlotov, S., Protopopov, M., Dzyuba, I., Vasylchuk, M., Panov, D. M., Tarkhanova, O., and Moroz, Y. Freedom Space 3.0: ML-assisted selection of synthetically accessible small molecules. *Journal of Chemical Information and Modeling*, 2025. doi: 10.1021/acs.jcim.5c01912.

Kim, H., Ryu, S., Jung, N., Yang, J., and Seok, C. CSearch: Chemical space search via virtual synthesis and global optimization. *Journal of Cheminformatics*, 16:137, 2024. doi: 10.1186/s13321-024-00936-8.

Kohonen, T. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990. doi: 10.1109/5.58325.

Liphardt, T. and Sander, T. Fast substructure search in combinatorial library spaces. *Journal of Chemical Information and Modeling*, 63(17):5133–5141, 2023. doi: 10.1021/acs.jcim.3c00290.

Liu, F., Mailhot, O., Glenn, I. S., Vigneron, S. F., Bassim, V., Xu, X., Fonseca-Valencia, K., Smith, M. S., Radchenko, D. S., Fraser, J. S., Moroz, Y. S., Irwin, J. J., and Shoichet, B. K. The impact of library size and scale of testing on virtual screening. *Nature Chemical Biology*, 21(7):1039–1045, 2025. doi: 10.1038/s41589-024-01797-w.

López Pérez, K., Jung, V., Chen, L., Huddleston, K., and Miranda-Quintana, R. A. BitBIRCH: Efficient clustering of large molecular libraries. *Digital Discovery*, 4:1042–1051, 2025. doi: 10.1039/D5DD00030K.

Lübbers, J., Schokolowski, M., Lessel, U., Weber, A., and Rarey, M. Guiding similarity search in chemical fragment spaces with weighted fingerprints. *Journal of Chemical Information and Modeling*, 66(4):2220–2229, 2026. doi: 10.1021/acs.jcim.5c02952.

Luo, S. and Coley, C. W. Efficient and programmable exploration of synthesizable chemical space, 2025.

McInnes, L., Healy, J., Saul, N., and Großberger, L. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

Medina-Franco, J. L., Martinez-Mayorga, K., Giulianotti, M. A., Houghten, R. A., and Pinilla, C. Visualization of the chemical space in drug discovery. *Current Computer-Aided Drug Design*, 4(4):322–333, 2008. doi: 10.2174/157340908786786010.

Narayan, A., Berger, B., and Cho, H. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology*, 39:765–774, 2021. doi: 10.1038/s41587-020-00801-7.

Patel, H., Ihlenfeldt, W. D., Judson, P. N., Moroz, Y. S., Pevzner, Y., Peach, M. L., Delannee, V., Tarasova, N. I., and Nicklaus, M. C. SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Scientific Data*, 7(1):384, 2020. doi: 10.1038/s41597-020-00727-4.

Pikalyova, R., Zabolotna, Y., Volochnyuk, D. M., and Varnek, A. Exploration of the chemical space of DNA-encoded libraries. *Molecular Informatics*, 41(6):e2100289, 2022. doi: 10.1002/minf.202100289.

Pikalyova, R., Zabolotna, Y., Horvath, D., Marcou, G., and Varnek, A. Chemical library space: Definition and DNA-encoded library comparison study case. *Journal of Chemical Information and Modeling*, 63(13):4042–4055, 2023a. doi: 10.1021/acs.jcim.3c00520.

Pikalyova, R., Zabolotna, Y., Horvath, D., Marcou, G., and Varnek, A. Meta-GTM: Visualization and analysis of the chemical library space. *Journal of Chemical Information and Modeling*, 63(17):5571–5582, 2023b. doi: 10.1021/acs.jcim.3c00719.

Pikalyova, R., Akhmetshin, T., Horvath, D., and Varnek, A. CoLiNN: A tool for fast chemical space visualization of combinatorial libraries without enumeration. *Molecular Informatics*, 44(3):e202400263, 2025. doi: 10.1002/minf.202400263.

Probst, D. and Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics*, 12:12, 2020. doi: 10.1186/s13321-020-0416-x.

Rarey, M. and Stahl, M. Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design*, 15(6):497–520, 2001. doi: 10.1023/A:1011144622059.

RDKit Developers. rdkit.Chem.rdSynthonSpaceSearch module, 2026. URL `https://www.rdkit.org/docs/source/rdkit.Chem.rdSynthonSpaceSearch.html`. Accessed 2026-05-05; RDKit 2025.09.6 documentation.

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

Sadybekov, A. V. et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature*, 601:452–459, 2022. doi: 10.1038/s41586-021-04220-9.

Schmidt, R., Klein, R., and Rarey, M. Maximum common substructure searching in combinatorial make-on-demand compound spaces. *Journal of Chemical Information and Modeling*, 62(9):2133–2150, 2022. doi: 10.1021/acs.jcim.1c00640.

Sosnin, S. Chemical space visual navigation in the era of deep learning and big data. *Drug Discovery Today*, 30(7):104392, 2025. doi: 10.1016/j.drudis.2025.104392.

Swanson, K. et al. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nature Machine Intelligence*, 6:338–353, 2024. doi: 10.1038/s42256-024-00809-7.

Tingle, B. I., Tang, K. G., Castanon, M., Gutierrez, J. J., Khurelbaatar, M., Dandarchuluun, C., Moroz, Y. S., and Irwin, J. J. ZINC-22: A free multi-billion-scale database of tangible compounds for ligand discovery. *Journal of Chemical Information and Modeling*, 63(4):1166–1176, 2023. doi: 10.1021/acs.jcim.2c01253.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `https://www.jmlr.org/papers/v9/vandermaaten08a.html`.

Walters, W. P. Virtual chemical libraries. *Journal of Medicinal Chemistry*, 62(3):1116–1124, 2019. doi: 10.1021/acs.jmedchem.8b01048.

Warr, W. A., Nicklaus, M. C., Nicolaou, C. A., and Rarey, M. Exploration of ultralarge compound collections for drug discovery. *Journal of Chemical Information and Modeling*, 62(9):2021–2034, 2022. doi: 10.1021/acs.jcim.2c00224.

Wetzel, S., Klein, K., Renner, S., Rauh, D., Oprea, T. I., Mutzel, P., and Waldmann, H. Interactive exploration of chemical space with Scaffold Hunter. *Nature Chemical Biology*, 5(8):581–583, 2009. doi: 10.1038/nchembio.187.

Zabolotna, Y., Lin, A., Horvath, D., Marcou, G., and Varnek, A. ChemSpace Atlas: Multiscale chemography of ultralarge libraries for drug discovery. *Journal of Chemical Information and Modeling*, 62(18):4537–4548, 2022a. doi: 10.1021/acs.jcim.2c00509.

Zabolotna, Y., Volochnyuk, D. M., Ryabukhin, S. V., Horvath, D., Gavrilenko, K. S., Marcou, G., Moroz, Y. S., Oksiuta, O., and Varnek, A. A close-up look at the chemical space of commercially available building blocks for medicinal chemistry. *Journal of Chemical Information and Modeling*, 62(9):2171–2185, 2022b. doi: 10.1021/acs.jcim.1c00811.

A. Additional Details

A.1. Additional Experimental Tables

Table 5. All-reaction ECFP4 exact-product retrieval on 42 queries from the 990,829-product slice. Overlap is exact top-100 overlap in percentage points; std. is over queries.

Policy	Mean	Median	Std.	Mean s
Reaction cap	63.4	67.0	17.1	1.13
Global synthons	59.2	63.5	16.9	0.80

Table 6. SYNTHONMAP diagnostics. Trustworthiness is computed on fixed 2,000-point subsamples against Jaccard neighborhoods of the 2048-bit OR fingerprints.

Map	Sample	Routes	Trust@15	Top sample	Top mass
Exact slice	60k	42/42	0.896	15.0%	39.4%
Freedom 4.0	1M	42/42	0.754	48.4%	77.7%

A.2. Proof of the Perturbation Bound

Let $V = \hat{U}$, $x_U = |Q \cap U|$, $x_V = |Q \cap V|$, $y_U = |Q \cup U|$, and $y_V = |Q \cup V|$. Changing U to V can change the query intersection by at most the number of flipped bits, so $|x_U - x_V| \leq |U \triangle V| = \delta$. The same is true for the query union, hence $|y_U - y_V| \leq \delta$. Since $x_U \leq y_U$ and $x_V \leq y_V$,

$$\left|\frac{x_U}{y_U} - \frac{x_V}{y_V}\right| = \frac{|x_U y_V - x_V y_U|}{y_U y_V} \tag{9}$$

$$\leq \frac{x_U|y_V - y_U| + y_U|x_U - x_V|}{y_U y_V} \tag{10}$$

$$\leq \frac{\delta}{y_V} + \frac{\delta}{y_V} \leq \frac{2\delta}{\min\{y_U, y_V\}}. \tag{11}$$

Thus the bound is informative when the OR/exact difference is small compared with the number of bits activated by the query-product union; it is loose when reaction execution creates or removes many query-relevant bits.

Importance weights for maps. If route r is sampled with probability p_r and then one synthon is sampled uniformly from each slot, a product from that route has proposal probability p_r/C_r. Product-uniform target mass is $1/C_{\text{tot}}$, so self-normalized density plots use weights proportional to C_r/p_r. The normalization constant cancels when densities are displayed after log transformation.

Exact-slice all-reaction check. In a 42-query all-reaction run on the same 990,829-product exact library, the keep-connector SYNTHONOR configuration recovered 63.4% mean exact top-100 overlap with mean wall time 1.13s/query under the reaction-cap policy, and 59.2% overlap with mean wall time 0.80s/query under the global-top-synthons policy. These numbers are consistent with the 10-query fingerprint-family snapshot in the main text and support the choice of ECFP4 with connector context as the default.

Figure provenance. The Freedom Space comparison and SYNTHONMAP paper figures were regenerated with repository R scripts. The SYNTHONMAP panels in Figure 2 are plotted in R from saved `synthonmap` sample and projection arrays generated from the Freedom Space 4.0 synthon table, cached ECFP4 synthon fingerprints, and the fitted 100,000-product UMAP anchor map. The camera-ready code repositories are `https://github.com/mireklzicar/synthonor` for SYNTHONOR and `https://github.com/mireklzicar/synthonmap` for SYNTHONMAP.

B. Limitations and Mitigations

Several approximation limits are direct consequences of using a binary OR surrogate. Because ordinary bit fingerprints do not store multiplicity, repeated local motifs collapse to the same representation as single occurrences; count Morgan fingerprints or capped-count composition are natural

mitigations when repeated environments matter. The surrogate also omits reaction-junction bits that appear only after assembly and can retain synthon bits that disappear during reaction execution. Junction correction features, optional product assembly for the final candidate set, and exact-product reranking are therefore important extensions for higher-stakes retrieval.

The search procedure can also inherit bit-saturation effects. Large or high-bit-density synthons may score well during prescreening because they cover many query bits, while final Tanimoto rescoring may later penalize their extra off-query bits. Bit-density calibration, Tversky-style query-containment scores, and explicit diversity or size controls could make this trade-off more predictable.

The map views have a separate visualization limitation: UMAP area is not conserved high-dimensional chemical density. Weighted density in these figures should be read as density in the learned display, not as a calibrated thermodynamic or chemical-space measure. Repeated-seed stability checks, densMAP comparisons, exact enumerable slices, and neighborhood-preservation diagnostics are the appropriate safeguards before interpreting fine map structure.