SynthFormer: Equivariant Pharmacophore-based Generation of Synthesizable Molecules for Ligand-Based Drug Design

Zygimantas Jocys[1],[*] **Zhanxing Zhu**[1], **Henriette M.G. Willems**[2], **Katayoun Farrahi**[1]

[1]University of Southampton, Southampton, United Kingdom
[2]The ALBORADA Drug Discovery Institute, University of Cambridge, Cambridge, United Kingdom

Abstract

Drug discovery is a complex process requiring significant time and cost to bring new medicines to patients. Many generative models aim to accelerate drug discovery, but few produce synthetically accessible molecules. Conversely, synthesis-focused models do not leverage the 3D information crucial for effective drug design. We introduce SynthFormer, a novel machine learning model that generates fully synthesizable molecules, structured as synthetic trees, by introducing 3D pharmacophores as input. SynthFormer features a 3D equivariant graph neural network to encode pharmacophores, followed by a Transformer-based synthesis-aware decoding mechanism for constructing synthetic trees as a sequence of tokens. It is a first-of-its-kind approach that could provide capabilities for designing active molecules based on pharmacophores, performing hit expansion and optimizing their properties. We demonstrate its effectiveness through various challenging tasks, including designing active compounds for a range of proteins, performing hit expansion and optimizing molecular properties.

1 Introduction

Drug discovery is a complex, lengthy, and expensive process (Wouters et al., 2020). Computer-aided drug design (CADD) has shown strong potential (Schneider and Fechner, 2005) with tools like Smina Koes et al. (2013) and Gold Jones et al. (1997), while generative machine learning (GML) offers a promising way to accelerate early discovery by efficiently exploring chemical space (Meyers et al., 2021). Two main computational approaches dominate early-stage discovery: *de novo* ligand design and virtual screening. Current GML methods, such as TargetDiff (Guan et al., 2023) and Pocket2Mol (Peng et al., 2022), often overlook synthesis feasibility. Pose prediction models like EquiBind (Stärk et al., 2022) and DiffDock (Corso et al., 2023) are restricted to fixed libraries and only predict poses.

In vitro screening approaches—DNA-encoded libraries (DELs) (Brenner and Lerner, 1992), high-throughput screening (HTS) (Inglese et al., 2007), and dose-response assays (Crump et al., 1976)—also face constraints. HTS is limited to fixed libraries, DELs by available reactions and building blocks, and dose-response assays by low throughput. Combining *in silico* and *in vitro* methods is thus essential to reduce costs and speed up discovery.

A key challenge for GML is synthesisability. Most models rely on the Synthetic Accessibility (SA) score (Ertl and Schuffenhauer, 2009), which poorly predicts true synthesis feasibility (Cretu et al., 2024; Gao and Coley, 2020). While some generative approaches build molecules from reactions (Gao et al., 2022; Cretu et al., 2024), none incorporate 3D structural information critical for modeling protein-ligand interactions. Pharmacophore-based ligand design remains essential when structures or binding poses are unreliable Dash et al. (2019). While AlphaFold Jumper (2021) and Boltz1

[*] Corresponding author: z.jocys@soton.ac.uk

Wohlwend et al. (2024) predict protein structures, they often lack local accuracy for structure-based design (Scardino et al., 2023; Wong et al., 2022), reinforcing the need for ligand-based approaches.

We introduce **SynthFormer**, a synthesis-aware model that translates 3D pharmacophores into molecules by decoding them as synthetic trees. Our main contributions:

- A novel model combining 3D pharmacophore information with synthesis constraints for realistic molecule generation.
- A synthesis-aware encoder-decoder that builds molecules as synthetic trees.
- Demonstration that SynthFormer produces diverse, novel compounds with strong docking performance and practical reaction pathways.

2 Related work

Combinatorial Optimization Combinatorial approaches build chemical space from reaction templates and building blocks, as in Galileo (Meyenburg et al., 2023), then explore it using optimization techniques like genetic algorithms (Gao et al., 2021) or Monte Carlo tree search (Swanson et al., 2023). Early tools like SYNOPSIS (Vinkers et al., 2003) generated molecules through virtual reactions and scoring. Recent deep learning advances predict reaction outcomes directly (Coley et al., 2019), reducing dependence on predefined templates.

Synthesis-Aware Deep Learning RL-based methods such as SynFlowNet Cretu et al. (2024) and SynNet Gao et al. (2022) sequentially optimize molecules while enforcing synthesizability, but rely on noisy reward functions like docking scores. Self-supervised approaches, like ChemProjector Luo et al. (2024), learn molecular representations by predicting transformations but cannot model binding interactions or activity cliffs Zhang et al. (2023). Fragment-based methods such as FLAG ZHANG et al. (2023), D3FG Lin et al. (2024), and JTVAE Jin et al. (2018) generate molecules by connecting fragments, though these connections are not synthesis-driven.

3D Generative Models Shape-based generation methods either condition on a protein structure or known molecules. Structure-based tools like TargetDiff (Guan et al., 2023), Pocket2Mol (Peng et al., 2022), DiffSBDD (Schneuing et al., 2024), and DiffBP (Lin et al., 2025) design molecules to fit binding sites but often fail to produce synthesis-ready compounds (Jocys et al., 2024; Luo et al., 2024). Ligand-based models like SQUID (Adams and Coley, 2023), LigDream (Skalic et al., 2019), and VoxMol (Pinheiro et al., 2024) encode ligands and decode new molecules. Pharmacophore-conditioned models also include PGMG (Zhu et al., 2022), TransPharmer (Xie et al., 2024), ShEPhERD (Adams, 2025), and PharmaDiff (Alakhadar, 2025).

Limitations of Existing Methods Combinatorial methods are slow and constrained by predefined templates. Deep learning models either focus on structural plausibility or synthesizability but rarely both, leading to outputs that lack practical feasibility, such as synthetic accessibility or assay compatibility. This gap between computational design and experimental reality limits their impact on real drug discovery pipelines.

3 Method

SynthFormer generates molecules from 3D pharmacophore features using an EGNN–transformer architecture with a synthesis-aware decoder (see Sec. 3.3), as illustrated in Figure 1. The model autoregressively predicts building blocks and reactions to produce coherent, interpretable molecules. To train and evaluate SynthFormer, we created a custom dataset. Existing datasets like PDBBind Wang et al. (2004), QM9 Wu et al. (2018), and QMugs Isert et al. (2022) focus on atom-level modeling, which is incompatible with our pharmacophore- and synthesis-tree-based approach. Leveraging Enamine's extensive chemical database, we built the first dataset designed for pharmacophore-driven synthetic tree generation, as detailed in the next section.

3.1 Encoder Input: Pharmacophore

We construct input batches by combining pharmacophore features and 3D spatial information from conformers. Pharmacophores are represented as multi-hot encoded vectors per atom, while conformers provide the corresponding (x, y, z) coordinates. Together, these form the EGNN encoder input, capturing both molecular geometry and pharmacophore profiles.

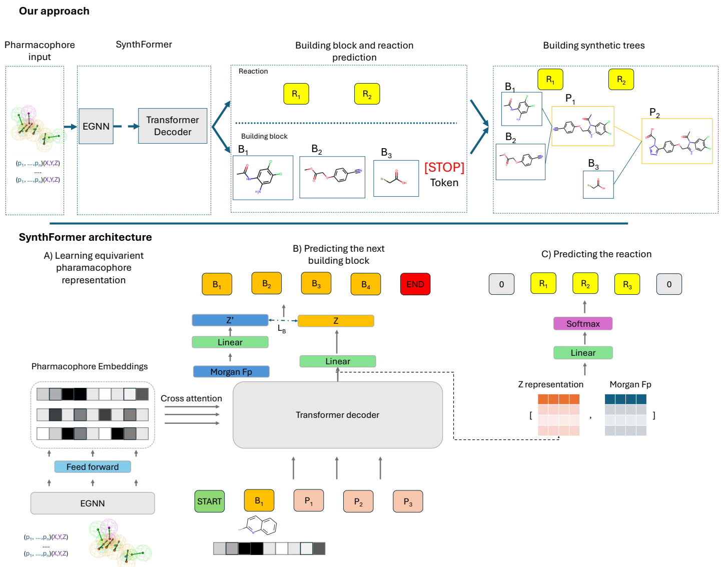


Figure 1: Pharmacophores are encoded as a fully connected graph and embedded with an EGNN. A transformer decoder then autoregressively predicts B_i and R_i from a start token, using fingerprints of previous B_i and positions P_i, until an end token is generated.

Conformers A conformer is a 3D arrangement of a molecule's atoms. Representing a molecule as a graph $G = (\mathcal{V}, \mathcal{E})$ with atoms $v \in \mathcal{V}$ and bonds $e \in \mathcal{E}$, its conformer space is $\mathcal{C}_G$. Each conformer $C \in \mathcal{C}_G$ is defined by intrinsic coordinates—bond lengths, bond angles, cycles—and torsion angles τ around rotatable bonds (Appendix A). For each atom v, the coordinates are given as $\mathbf{x}_v^k = (x_v^k, y_v^k, z_v^k)$ with $k \in \{1, \dots, n_v\}$, where n_v is the number of valid conformations.

Pharmacophores We use six RDKit (Landrum, 2006)-supported pharmacophore types: Hydrogen Bond Donors (HBD), Hydrogen Bond Acceptors (HBA), Aromatic Rings (AR), Hydrophobic Centers (HC), Positive Ionizable features (PIF), and Negative Ionizable features (NIF). Each atom's pharmacophore vector is encoded as $\mathbf{p} = [p_1, \dots, p_n]$, where $p_i \in \{0, 1\}$ indicates feature presence.

Data Preparation We generate data by randomly selecting building blocks and performing a sequence of n random reactions, storing the resulting products. This process is repeated, sampling both initial molecules and previously generated products. For each sampled molecule, we generate a random conformer using MMFF optimization in RDKit, then extract its pharmacophores and 3D coordinates. These form the model inputs, while the target output is the molecule's synthetic tree.

3.2 Decoder Output: Molecules as Synthetic Trees

The generation begins with building blocks and proceeds by iteratively applying reactions. For simplicity, we assume each reaction produces a single product. As shown in Figure 1, each node in the tree represents a compound, and each edge represents a reaction transforming one compound into another. This structure enables predicting feasible synthetic routes, optimizing pathways, and planning step-by-step molecule construction. For example, starting with building block B_1, reaction R_1 combines it with B_2 to form product P_1. Then, reaction R_2 uses P_1 and B_3 to create P_3. This process repeats, with each product feeding into the next reaction, ultimately forming the target molecule as a complete synthetic tree. When sampling batches, we select a target molecule and trace backward through all preceding steps, enabling efficient exploration of possible synthesis paths.

3.3 Architecture

Equivariant Pharmacophore Encoder We encode pharmacophoric profiles using Cartesian coordinates in a fully connected graph. The EGNN processes node embeddings $\mathbf{p}$, coordinates $\mathbf{x}'$, and edges $\mathcal{E}$:

$$m_{ij} = \phi_e\left(\mathbf{h}_i^l, \mathbf{h}_j^l, \left\|\mathbf{x}_i' - \mathbf{x}_j'\right\|_2^2, e_{ij}\right) \tag{1}$$

$$\mathbf{x}_i'^{l+1} = \mathbf{x}_i' + \sum_{j\neq i}(\mathbf{x}_i' - \mathbf{x}_j')\phi_x(m_{ij}) \tag{2}$$

$$m_i = \sum_{j\in\mathcal{N}(i)} m_{ij} \tag{3}$$

$$\mathbf{h}_i^{l+1} = \phi_h\left(\mathbf{h}_i^l, m_i\right) \tag{4}$$

where ϕ_e, ϕ_x, and ϕ_h are MLPs. Seven EGNN layers produce embeddings for cross-attention.

PDBID	**Ref Dock** (kcal/mol)	**Av. Dock Gen (Min)** (kcal/mol)↓	**Random Baseline** (kcal/mol)	**Murcko**↓	**Tanimoto**↓	**P4 Sim**↑
1x8d	-6.17	-6.43 (-8.33)	-4.62	0.07	0.10	0.71
1xbo	-10.79	-8.22 (-11.22)	-3.75	0.11	0.12	0.48
2afw	-3.85	-7.92 (-9.21)	-4.96	0.08	0.09	0.67
2aog	-10.45	-8.63 (-11.53)	-4.13	0.01	0.07	0.43
2bt9	-6.68	-7.63 (-10.11)	-5.72	0.03	0.08	0.71
3coy	-12.41	-10.14 (-12.26)	-4.08	0.07	0.08	0.53
3ga5	-9.68	-4.55 (-7.60)	-4.62	0.04	0.05	0.58
4q6d	-7.41	-7.55 (-10.07)	-3.52	0.10	0.07	0.82
5fl4	-8.09	-8.01 (-9.44)	-3.94	0.12	0.11	0.57
5ka1	-7.67	-6.44 (-9.71)	-5.21	0.07	0.07	0.70

Table 1: Docking energies (kcal/mol) for reference ligands (Ref), generated molecules (Gen, min in brackets), and random baseline (Rand). Murcko, Tanimoto, and P4 measure similarity to Ref ligands.

Synthesis-Aware Decoder The transformer processes Morgan fingerprints $f_p \in \{0,1\}^{4096}$ with radius 3, starting with $[START]$ and ending with $[END]$. Standard multihead attention with $N = 7$ layers and $h = 8$ heads generates representation Z:

$$B_{i+1} = argmax(\text{Similarity}\,(Z_i, Z'))) \tag{5}$$

$$R_{i+1} \sim \text{Softmax}(\text{Linear}(\text{Concat}(Z_i, \text{f}_p(B_{i+1})))) \tag{6}$$

3.4 Training

Causal masking ensures autoregressive generation. The building block loss uses cosine similarity:

$$L_{\text{B}} = \frac{1}{n}\sum_{j=0}^{l-1} \frac{Z'_{i+1} \cdot Z_{i+1}}{||Z'_{i+1}|| \cdot ||Z_{i+1}||} \tag{7}$$

where $Z' = \mathbf{W} \cdot \text{f}_p(B_{i+1}) + \mathbf{b}$. The reaction loss uses cross-entropy:

$$L_{\text{rxn}} = \frac{1}{m}\sum_{i=0}^{\ell-1} \text{CE}(\hat{r}_{i+1}, r_{i+1}) \tag{8}$$

$$L = L_{\text{B}} + L_{\text{rxn}} \tag{9}$$

3.5 Inference

Starting with pharmacophore data and $[START]$, the decoder autoregressively predicts building blocks B_i and reactions R_i, generating products until $[END]$ terminates the process.

4 Experiments

We evaluate our model's active ligand design performance against random baselines and synthesis-aware models, comparing with three ligand-based 3D models on docking and synthesizability. We explore embedding utility for analog generation and molecular optimization.

4.1 Experimental Setup

We use **58 publicly available reaction templates** in SMARTS format (Hartenfeller et al., 2012) and **251,222 commercially available building blocks** from Enamine Ltd. (2025). To compare molecules, we compute **similarity scores** using Tanimoto similarity across three fingerprints: the Morgan fingerprint (length 4096, radius 2) (Morgan, 1965), the Morgan fingerprint of the Murcko scaffold (Bemis and Murcko, 1996), and pharmacophore similarity via RDKit, with all scores normalized to the ([0, 1]) range. For structure-based evaluation, we perform **docking** using SMINA Koes et al. (2013), converting structures to PDBQT format with OpenBabel O'Boyle et al. (2011), and using a 25×25×25 Å docking box centered on the ligand centroid to generate up to 10 poses.

4.2 Main Evaluation: Designing Active Compounds

Docking scores were compared with original ligands to evaluate binding affinity and pose. For PDB entries 1x8d, 2bt9, 416d, and 5fl4, docking energy closely matched the reference, with lowest docking energies consistently outperforming reference, except for 3ga5 (Table 1). Docking energies for generated molecules ranged from -4.55 to -10.14 kcal/mol (Appendix A, Figure 3).

Similarity metrics Tanimoto and Murcko indicate extremely low structural resemblance between generated molecules and reference ligands, ranging from 0.06 to 0.12. Generated molecules adhere to the Ghose filter with molecular weights 305.96-339.02 g/mol and LogP values within acceptable range (-0.4 to 5.6). Pharmacophore similarity scores range 0.53-0.81, containing most pharmacophores on average. The low similarity score of 0.43 for 2AOG results from the large reference ligand size (772 Da), creating feature alignment mismatch with smaller compounds.

Method	**Type**	**Δ Dock $\downarrow$**	**Tanimoto $\downarrow$**	**Synthesis $\uparrow$**
ChemProjector	Synthesis	2.55	0.53	100%
SynNet	Synthesis	2.7	0.64	100%
SQUID	3D	**2.39**	0.24	23.6%
Ligdream	3D	2.78	0.22	32.4%
SynthFormer	Synthesis+3D	2.46	**0.09**	100%

Table 2: Comparison of synthesis-aware 3D generative methods and SynthFormer across 10 PDBs (100 compounds each). SynthFormer shows lowest similarity to reference ligands while maintaining docking performance comparable to models

Comparison to Random Baseline Best generated molecules with optimal poses consistently outperform reference molecules in docking energy. Generated molecules consistently beat the random baseline (100 molecules sampled from Enamine building blocks and available reactions docked to binding sites), indicating robust effectiveness in generating molecules with enhanced docking properties.

Comparison Against Existing Methods Table 2 presents an analysis for generating 100 molecules per target. ChemProjector and SynNet are inherently designed for molecule-to-molecule tasks, naturally generating analogs. Synthesis scores for 3D models were calculated using Drug Pose Jocys et al. (2024) with Enamine Real Space.

SynthFormer achieves Δ Dock score of 2.46, comparable to SQUID (2.39) and better than Ligdream (2.78). For chemical diversity, SynthFormer significantly outperforms with Tanimoto score 0.09 vs. SQUID (0.24) and Ligdream (0.22), indicating more distinct molecules from known compounds. SynthFormer achieves 100% synthesizability vs. SQUID (23.6%) and Ligdream (32.4%), demonstrating all molecules are synthetically feasible.

4.3 Other Evaluations

Additional SynthFormer capabilities for hit expansion and compound optimization, with evidence of meaningful encoding space organization. No existing benchmarks exist for this framework.

Building Block Encoding Exploration Building block encodings from transformer model were extracted and analyzed using cosine similarity to quantify structural closeness. Analysis revealed high similarity in preserving ring structures and nitrogen atom counts, with encodings capturing these structural features with remarkable fidelity.

Hit Expansion The model identifies synthesizable analogs of hit molecules (Keseru and Makara, 2006; Levin et al., 2023). Crystallized ligands are encoded as Morgan fingerprints, combined with a [Start] token to sample building blocks and reactions. 100 analogs per seed (Appendix A Figure 4) are docked, and top poses selected. 11.4% outperform the original hits while maintaining a mean Tanimoto similarity of 0.67.

Molecule Optimization Implemented a genetic algorithm (Appendix B) for molecular optimization within a synthetic reaction tree. For 10 PDB molecules, reagents were adjusted using cosine similarity to generate new molecules and resample reactions. Over three cycles, logP decreased by 0.21, druglikeness increased by 0.09, and average energy increased by 0.03 kcal/mol.

5 Conclusions

SynthFormer introduces a method for generating synthetically accessible molecules from pharmacophores. By combining a 3D equivariant GNN with a synthesis-aware decoder, it designs both active and synthesizable molecules. SynthFormer outperforms random baselines in docking and shows higher synthesizability than existing 3D models, supporting its potential to accelerate drug discovery, hit expansion, and molecule optimization.

References

Adams, e. a. (2025). Diffusing shape, electrostatics, and pharmacophores for small-molecule design. *arXiv preprint arXiv:2411.04130*.

Adams, K. and Coley, C. W. (2023). Equivariant shape-conditioned generation of 3d molecules for ligand-based drug design. In *The Eleventh International Conference on Learning Representations*.

Alakhadar, e. a. (2025). Pharmacophore-conditioned diffusion model for ligand generation. *arXiv preprint arXiv:2505.10545*.

Bemis, G. W. and Murcko, M. A. (1996). The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893. PMID: 8709122.

Brenner, S. and Lerner, R. A. (1992). Encoded combinatorial chemistry. *Proceedings of the National Academy of Sciences of the United States of America*, 89(12):5381–5383.

Coley, C. W., Jin, W., Rogers, L., Jamison, T. F., Jaakkola, T. S., Green, W. H., Barzilay, R., and Jensen, K. F. (2019). A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science*, 10(2):370–377.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. (2023). Diffdock: Diffusion steps, twists, and turns for molecular docking.

Cretu, M., Harris, C., Roy, J., Bengio, E., and Liò, P. (2024). Synflownet: Towards molecule design with guaranteed synthesis pathways.

Crump, K. S., Hoel, D. G., Langley, C. H., and Peto, R. (1976). Fundamental carcinogenic processes and their implications for low dose risk assessment. *Cancer Research*, 36(9_Part_1):2973–2979.

Dash, R. C., Ozen, Z., McCarthy, K. R., Chatterjee, N., Harris, C. A., Rizzo, A. A., Walker, G. C., Korzhnev, D. M., and Hadden, M. K. (2019). *ChemMedChem*, 14:1610.

Ertl, P. and Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1.

Gao, W. and Coley, C. W. (2020). The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.*, 60:5714–5723.

Gao, W., Mercado, R., and Coley, C. W. (2021). Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. In *International Conference on Learning Representations*.

Gao, W., Mercado, R., and Coley, C. W. (2022). Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design.

Guan, J., Qian, W. W., Peng, X., Su, Y., Peng, J., and Ma, J. (2023). 3d equivariant diffusion for target-aware molecule generation and affinity prediction.

Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., and Schneider, G. (2012). Dogs: Reaction-driven de novo design of bioactive compounds. *PLOS Computational Biology*, 8(2):1–12.

Inglese, J., Shamu, C., and Guy, R. K. (2007). Reporting data from high-throughput screening of small-molecule libraries. *Nature Chemical Biology*, 3(8):438–441.

Isert, C., Atz, K., Jiménez-Luna, J., et al. (2022). Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):273.

Jin, W., Barzilay, R., and Jaakkola, T. S. (2018). Junction tree variational autoencoder for molecular graph generation. *CoRR*, abs/1802.04364.

Jocys, Z., Grundy, J., and Farrahi, K. (2024). Drugpose: benchmarking 3d generative methods for early stage drug discovery. *Digital Discovery*, 3:1308–1318.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267(3):727–748.

Jumper, J. e. a. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589.

Keseru, G. M. and Makara, G. M. (2006). Hit discovery and hit-to-lead approaches. *Drug Discovery Today*, 11(15-16):741–748.

Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013). Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904.

Landrum, G. (2006). Rdkit: Open-source cheminformatics software. *http://www.rdkit.org*.

Levin, I., Fortunato, M. E., Tan, K. L., and Coley, C. W. (2023). Computer-aided evaluation and exploration of chemical spaces constrained by reaction pathways. *AIChE Journal*, 69(12):e18234.

Lin, H., Huang, Y., Zhang, O., Ma, S., Liu, M., Li, X., Wu, L., Wang, J., Hou, T., and Li, S. Z. (2025). Diffbp: generative diffusion of 3d molecules for target protein binding. *Chem. Sci.*, 16:1417–1431.

Lin, H., Huang, Y., Zhang, O., Wu, L., Li, S., Chen, Z., and Li, S. Z. (2024). Functional-group-based diffusion for pocket-specific molecule generation and elaboration.

Ltd., E. (2025). Building blocks & screening compounds. `https://enamine.net`. Accessed: 2025-01-24.

Luo, S., Gao, W., Wu, Z., Peng, J., Coley, C. W., and Ma, J. (2024). Projecting molecules into synthesizable chemical spaces.

Meyenburg, C., Dolfus, U., Briem, H., et al. (2023). Galileo: Three-dimensional searching in large combinatorial fragment spaces on the example of pharmacophores. *Journal of Computer-Aided Molecular Design*, 37:1–16.

Meyers, J., Fabian, B., and Brown, N. (2021). De novo molecular design and generative models. *Drug Discovery Today*, 26(11):2707–2715.

Morgan, H. L. (1965). The generation of a unique chemical fingerprint. *Journal of Chemical Documentation*, 5(2):107–120.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33.

Peng, X., Luo, S., Guan, J., Xie, Q., Peng, J., and Ma, J. (2022). Pocket2Mol: Efficient molecular sampling based on 3D protein pockets. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17644–17655. PMLR.

Pinheiro, P. O., Rackers, J., Kleinhenz, J., Maser, M., Mahmood, O., Watkins, A. M., Ra, S., Sresht, V., and Saremi, S. (2024). 3d molecule generation by denoising voxel grids.

Scardino, A., D'Alonzo, D., Cuzzolin, A., Sturlese, M., and Moro, S. (2023). How good are alphafold models for docking-based virtual screening? *iScience*, 26(1):105920.

Schneider, G. and Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4:649–663.

Schneuing, A., Harris, C., Du, Y., Didi, K., Jamasb, A., Igashov, I., Du, W., Gomes, C., Blundell, T., Lio, P., Welling, M., Bronstein, M., and Correia, B. (2024). Structure-based drug design with equivariant diffusion models.

Skalic, M., Jiménez, J., Sabbadin, D., and De Fabritiis, G. (2019). Shape-based generative modeling for de novo drug design. *Journal of Chemical Information and Modeling*, 59(3):1205–1214. PMID: 30762364.

Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R., and Jaakkola, T. (2022). Equibind: Geometric deep learning for drug binding structure prediction.

Swanson, K., Liu, G., Catacutan, D., Zou, J., and Stokes, J. (2023). Generative ai for designing and validating easily synthesizable and structurally novel antibiotics. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*.

Vinkers, H. M., de Jonge, M. R., Daeyaert, F. F., Heeres, J., Koymans, L. M., van Lenthe, J. H., Lewi, P. J., Timmerman, H., Van Aken, K., and Janssen, P. A. (2003). Synopsis: Synthesize and optimize system in silico. *Journal of Medicinal Chemistry*, 46(13):2765–2773.

Wang, R., Fang, X., Lu, Y., Yang, C. Y., and Wang, S. (2004). The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980.

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. (2024). Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*.

Wong, F., Krishnan, A., Zheng, E. J., Stokes, J. M., Keiser, M. J., Andrews, I. W., Cherry, J. M., and Brown, E. D. (2022). Benchmarking alphafold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology*, 18(9):e11081.

Wouters, O. J., McKee, M., and Luyten, J. (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844–853.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). Moleculenet: A benchmark for molecular machine learning.

Xie, W., Zhang, J., Xie, Q., Gong, C., Xu, Y., Lai, L., and Pei, J. (2024). Accelerating discovery of novel and bioactive ligands with pharmacophore-informed generative models. *arXiv preprint arXiv:2401.01059*.

ZHANG, Z., Min, Y., Zheng, S., and Liu, Q. (2023). Molecule generation for target protein binding with structural motifs. In *The Eleventh International Conference on Learning Representations*.

Zhang, Z., Zhao, B., Xie, A., Bian, Y., and Zhou, S. (2023). Activity cliff prediction: Dataset and benchmark.

Zhu, H., Zhou, R., Tang, J., and Li, M. (2022). Pgmg: A pharmacophore-guided deep learning approach for bioactive molecular generation. *arXiv preprint arXiv:2207.00821*.

6 Appendix

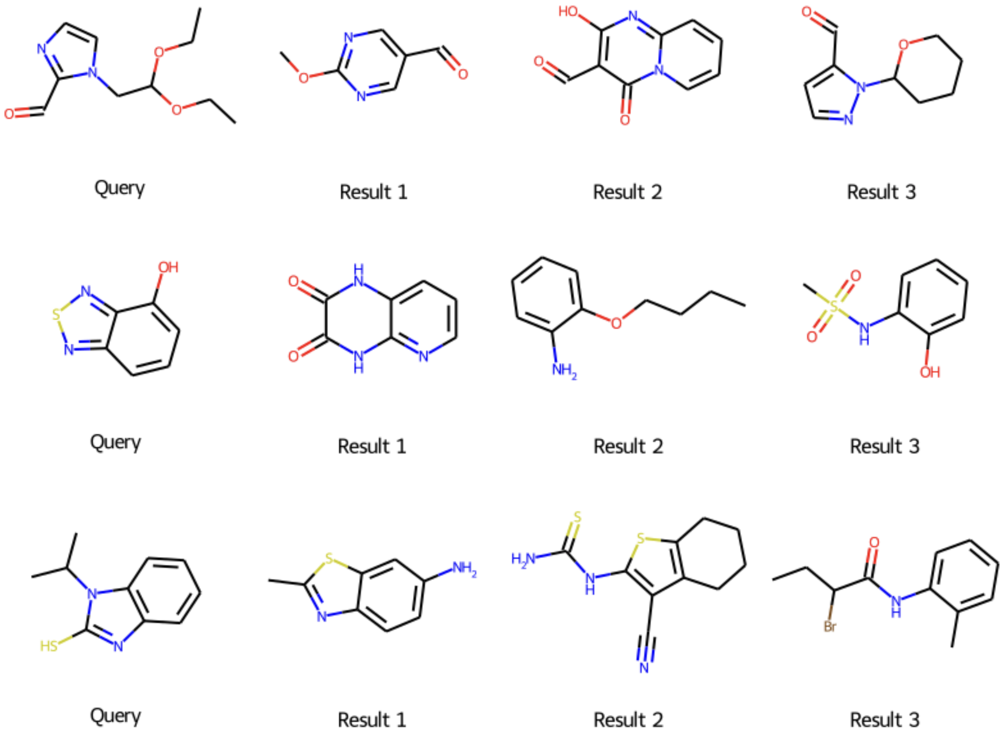
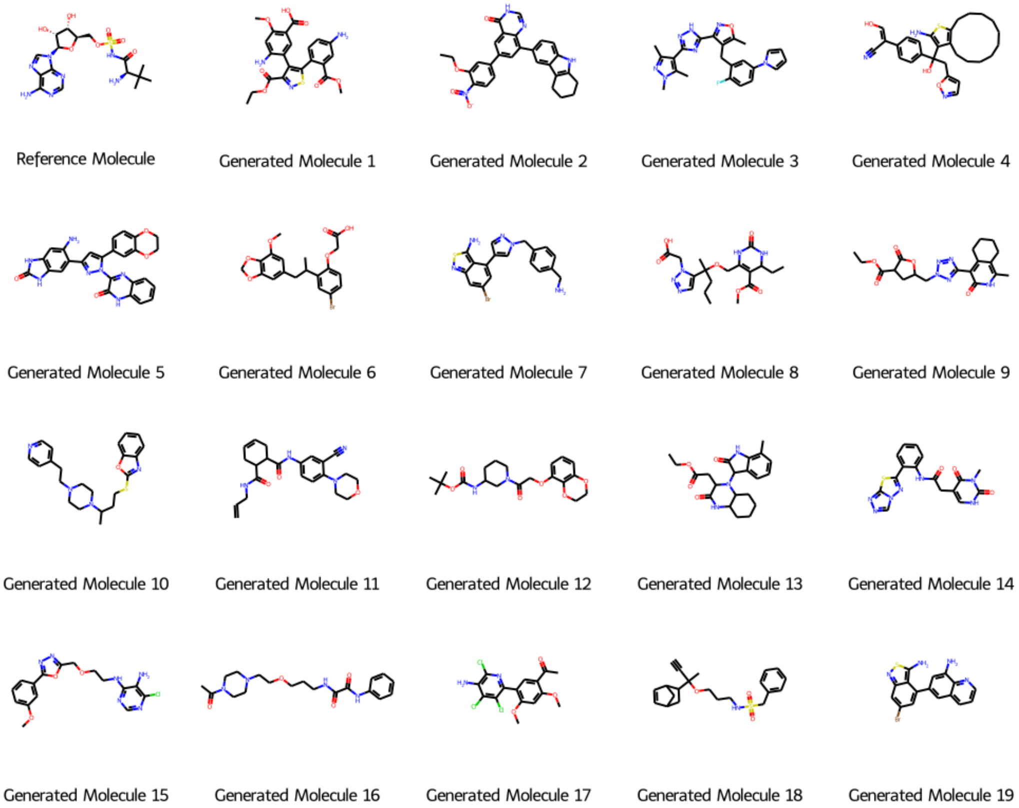


Figure 2: The building block encoding of the query molecule serves as the reference, with the three closest molecules identified using cosine similarity, preserving significant structural similarity.



Reference Molecule

Generated Molecule 1

Generated Molecule 2

Generated Molecule 3

Generated Molecule 4

Generated Molecule 5

Generated Molecule 6

Generated Molecule 7

Generated Molecule 8

Generated Molecule 9

Generated Molecule 10

Generated Molecule 11

Generated Molecule 12

Generated Molecule 13

Generated Molecule 14

Generated Molecule 15

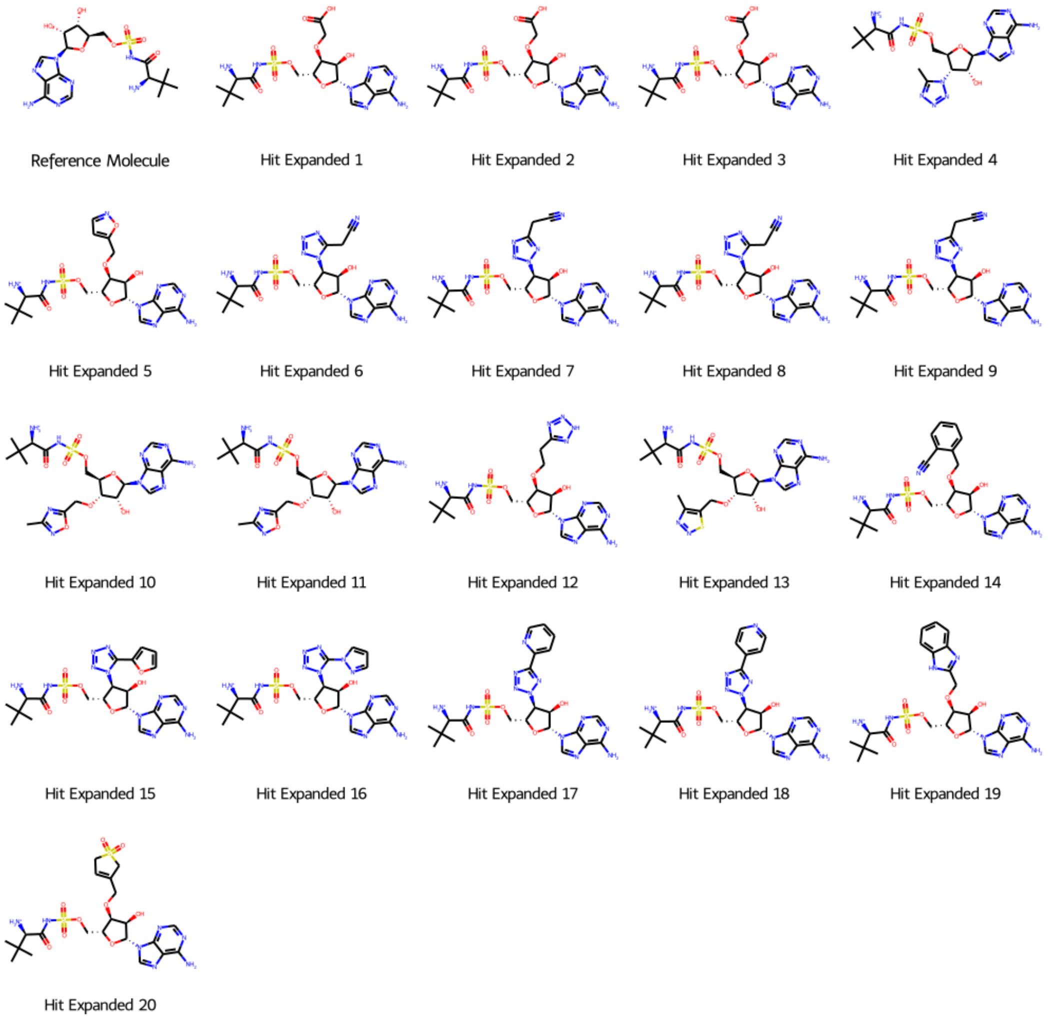
Generated Molecule 16

Generated Molecule 17

Generated Molecule 18

Generated Molecule 19

Figure 3: The first molecule corresponds to the reference structure derived from the 3COY PDB ID. The subsequent molecules are computationally generated using the SynthFormer model, illustrating its capability to design novel molecular structures based on a known protein-ligand complex. These results highlight SynthFormer's potential in generating diverse and plausible molecular candidates.



Reference Molecule

Hit Expanded 1

Hit Expanded 2

Hit Expanded 3

Hit Expanded 4

Hit Expanded 5

Hit Expanded 6

Hit Expanded 7

Hit Expanded 8

Hit Expanded 9

Hit Expanded 10

Hit Expanded 11

Hit Expanded 12

Hit Expanded 13

Hit Expanded 14

Hit Expanded 15

Hit Expanded 16

Hit Expanded 17

Hit Expanded 18

Hit Expanded 19

Hit Expanded 20

Figure 4: The first molecule corresponds to the reference structure derived from the 3COY PDB ID. The subsequent molecules are expanded hits generated by the SynthFormer model, demonstrating structural diversity and novel chemical scaffolds.