Synthesizable by Design: A Retrosynthesis-Guided Framework for Molecular Analog Generation

Shuan Chen[a,b], Gunwook Nam[a], and Yousung Jung*[a,b,c]

[a]   Department of Chemical and Biological Engineering (BK21 four)
        Seoul National University
        1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

[b]   Institute of Chemical Processes
        Seoul National University
        1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

[c]   Interdisciplinary Program in Artificial Intelligence
        Seoul National University
        1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

*Email: yousung.jung@snu.ac.kr

Abstract: The disconnect between AI-generated molecules with desirable properties and their synthetic feasibility remains a critical bottleneck in computational drug and material discovery. While generative AI has accelerated the proposal of candidate molecules, many of these structures prove challenging or impossible to synthesize using established chemical reactions. Here, we introduce SynTwins, a novel retrosynthesis-guided molecular analog design framework that designs synthetically accessible molecular analogs by emulating expert chemist strategies through a three-step process: retrosynthesis, similar building block searching, and virtual synthesis. In comparative evaluations, SynTwins demonstrates superior performance in generating synthetically accessible analogs compared to state-of-the-art machine learning models while maintaining high structural similarity to original target molecules. Furthermore, when integrated with existing molecule optimization frameworks, our hybrid approach produces synthetically feasible molecules with property profiles comparable to unconstrained molecule generators, yet its synthesizability ensured. Our comprehensive benchmarking across diverse molecular datasets demonstrates that SynTwins effectively bridges the gap between computational design and experimental synthesis, providing a practical solution for accelerating the discovery of synthesizable molecules with desired properties for a wide range of applications

Introduction

The discovery of novel molecules with specific chemical properties is a critical yet challenging process in pharmaceutical and chemical industries, often requiring years or decades due to the vastness of chemical space[1,2]. Generative artificial intelligence (AI) has emerged as a powerful accelerator for this process, rapidly proposing candidate molecules with target properties[3,4]. When paired with accurate property prediction models, these AI approaches enable efficient computational screening of molecules against desired criteria. However, a fundamental limitation persists: a significant portion of AI-generated molecules are difficult or impossible to synthesize using known chemical reactions and available building blocks[5]. While synthetic accessibility scoring can partially guide generative models toward more feasible structures, many proposed molecules remain synthetically challenging, creating a disconnect between computational design and experimental implementation.

To address this synthesis planning bottleneck, computational tools for single-step and multi-step retrosynthesis have been developed[6–8]. However, even with these tools, many AI-generated molecules remain synthetically challenging. In such cases, generating structurally similar synthetically accessible analogs offers a promising alternative to bypass synthesis difficulties while preserving desired properties, similar to how medicinal chemists have designed accessible analogs of complex natural products[9,10]. In real-world laboratory settings, synthesis capabilities vary considerably—pharmaceutical labs frequently rely on amide formations and heterocycle synthesis, while materials science teams favor reactions like Suzuki coupling for developing OLEDs. Additionally, the growing emphasis on sustainable chemistry has shifted preferences toward greener reactions and building blocks[11–13]. Given these constraints, the set of synthetically accessible molecules varies across different research environments. Consequently, researchers are exploring strategies for designing molecules by virtually synthesizing them from predefined reaction sets and available building block libraries.

Recent advances have applied machine learning (ML) to address these challenges through two main strategies. The first approach focuses on explicitly generating synthesizable molecules by using reinforcement learning[14] and GFlowNet-based methods[15,16] to select reactants and reaction pathways optimized for both synthetic feasibility and target properties (**Figure 1b**). The second strategy designs synthetically accessible analogs, structurally similar alternatives to promising but challenging-to-synthesize AI-generated molecules (**Figure 1c**). For example, Noh et al.[17] created a variational autoencoder (VAE) that encodes multi-step synthesis sequences into a latent space to generate complex yet synthetically feasible molecules. Similarly, Gao et al.[18] formulated the problem as a Markov decision process (MDP) with an amortized solution approach. While these methods represent promising directions, existing

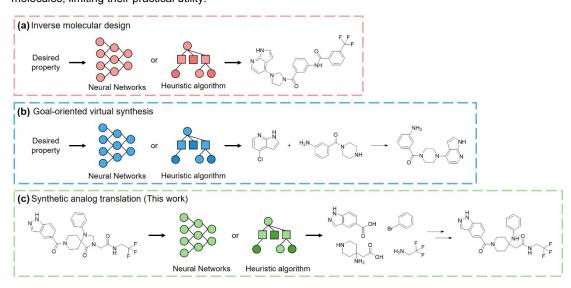models typically achieve suboptimal structural similarity to target molecules, limiting their practical utility.

Figure 1. The three strategies for designing novel molecules using computational methods. (a) Inverse molecular design model. (b) Conditional virtual synthesis model. (c) Synthetically accessible analog design.

We introduce SynTwins, a novel tree search algorithm that designs synthetically accessible molecular analogs by emulating the intuitive strategies employed by experienced medicinal and synthetic chemists. SynTwins implements a three-step process: (1) retrosynthetic analysis of target molecules to identify key structural components, (2) systematic searching for similar yet readily available building blocks that maintain essential pharmacophores, and (3) virtual synthesis using well-established reaction templates to assemble the final molecular analog. Our comprehensive evaluation demonstrates that SynTwins, a non-ML-based approach, outperforms state-of-the-art ML-based models in generating synthetically accessible analogs preserving high structural similarity to original target molecules. Furthermore, by integrating SynTwins with existing ML-based molecule optimization pipelines, our hybrid approach produces synthetically feasible molecules with comparable bioactivity and physicochemical profiles to those generated by unconstrained molecular optimizers. This work presents a practical solution to the synthesis-design gap and establishes a foundation for more effective molecular discovery pipelines that successfully translate computational designs into laboratory syntheses.

Results and Discussion

SynTwins

SynTwins is a synthetically accessible analog search algorithm that mirrors the intuitive workflow of chemists in a laboratory. Typically, chemists first evaluate whether a target molecule can be synthesized directly using available reactions and building blocks. If direct synthesis is not feasible due to the unavailability of specific precursors, they explore alternative building blocks with structural similarities to construct analogous molecule. Following this intuition, SynTwins systematically explores

potential analogs of a target molecule through a combined iterative process of retrosynthesis analysis and virtual synthesis. The algorithm operates in three distinct phases: retrosynthesis, building block search, and virtual synthesis. Computationally, the full processes are achieved by the following three phases of SynTwins (**Figure 2**):

1. Retrosynthesis: The first step of SynTwins is to perform multi-step retrosynthesis using retro-reaction templates[19] (denoted as T_r, detailed in the next subsection) to build a synthesis tree. For retrosynthesis, we apply all the compatible retro-reaction templates to the target molecule at each retrosynthesis step instead of applying ML-based retrosynthesis models in this work. The synthesis tree is expanded until maximum tree depth (d_{max}) is reached. The default number of d_{max} is 3 in this study. The target molecule and the precursors found in the synthesis tree are denoted as P_{ref} and $\{R_{ref}\}$, where $\{R_{ref}\}$ may or may not include the ones in the set of available building blocks. More details about the retrosynthesis are given in Method section.

2. Building block search: After the retrosynthesis is finished, the building blocks that are structurally similar to the found precursors $\{R_{ref}\}$ are searched using an k-nearest neighbor (kNN) algorithm[20], which search for the molecules that have the shortest distance in the chemical space according to their Extended Connectivity Fingerprints (ECFP)[21]. The building blocks found by the kNN algorithms are denoted as $\{R_{twin}\}$, and these molecules are limited to have the same functional groups presented in T_r to guarantee the compatibility of applying the reverse reaction templates used for the virtual synthesis in the next step. The default number of neighbors (k) of kNN search is 10 in this study. More details about the kNN implementation are given in Method section.

3. Virtual synthesis: Finally, k^2 molecule analogs $\{P_{twin}\}$ are synthesized by applying the forward-reaction template (T_f), which represent the reverse version of the retro-reaction template T_r used in the first phase, on every pair of building block searched in the last phase $\{R_{twin}\}$.
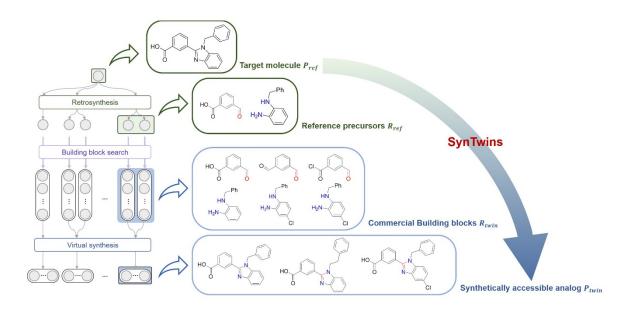
Figure 2. An example of generating synthetically accessible analogs using SynTwins. The imidazole in the target molecule P_{ref} is decoupled to two precursors R_{ref} containing an aldehyde and diamine functional groups, respectively. Next, the building blocks structurally similar to the precursors R_{twin} were searched from the list of available molecules. Finally, the synthetically accessible analogs P_{twin} are virtually synthesized by the same imidazole synthesis reaction using the aldehyde and diamine functional groups in the building blocks.

Retro-reaction templates

The chemical reactions available in this study are represented by forward-reaction templates in the form of Simplified Molecular Input Line Entry System of Reactions (SMIRKS), where the reactive substructure patterns are described in the SMILES arbitrary target specification (SMARTS) format[22]. Since SMARTS-based forward-reaction templates can define multiple interchangeable functional groups within a single template, one forward-reaction template can cover various functional group combinations in the same type of reaction. For example, as the halide used in the Suzuki Coupling is simplified by "X", which accepts either Cl, Br, I to run the reaction, it makes direct derivation of retro-reaction templates by simply reversing reactants and products infeasible. To address this, we derive retro-reaction templates by extracting them from 10,000 virtually synthesized reactions generated from the available building blocks and forward-reaction templates using RDChiral[23] (See he details in **Supporting Information**).

To balance the specificity and generalizability of retro-reaction templates, each retro-reaction template only includes the atoms that are one bond away from the changing reaction center, without explicitly defining functional groups. This representation is similar to the extended local reaction template (ELRT) used in LocalMapper[24]. Consequently, the number of retro-reaction templates is approximately 10 times greater than the number of forward-reaction templates in our experiments.

Synthesizable accessible analog design

To run SynTwins for our experiments, we followed existing works[14,17,18,25,26] and collected 58 reaction templates from Hartenfeller et al.[27] and 64 reaction templates from Button et al.[28]. After removing the duplications, where two reaction templates from different literatures represent the same reaction, we removed 21 reaction templates and kept the remaining 101 forward-reaction templates as our reaction collections. For the building blocks used in this study, we use a set of 150,560 commercially available molecules from the Enamine Building Block catalogue Global Stock[29] used in the previous study[17,30]. With these building blocks and forward-reaction templates, we derived 1,163 retro-reaction templates for retrosynthesis. More details about the selected reaction templates are given in the **Supporting Information.**

To evaluate the performance of SynTwins on generating the synthetically accessible analogs, we compare our method with two state-of-the-art ML-based models, ChemProjector[25] and SynFormer[26], after training the models again with the same set of reaction templates and building blocks. We examine the performance of these algorithms using 1,000 virtually synthesized products, 1,000 molecules from ChEMBL[31], 170 molecules from US Patent Trade Office (USPTO) dataset[32], and 100 molecules from FDA-approved drugs[33]. The details of curating molecules are given in the Method section.

The results are evaluated by reconstruction rate and top-k average similarity. The reconstruction rate shows the percentage of generated molecule analogs being exactly same with the target molecules, and the top-k average similarity is a metric to measure how structurally similar are the top-k generated molecule analogs are compared to the target molecule. The structural similarity is calculated by the Tanimoto similarity[34,35] of 4096-bits ECFP4[21].

As shown in **Table 1**, SynTwins consistently outperforms the baseline methods across all datasets in both reconstruction capability and molecular similarity. Compared to ChemProjector and SynFormer, SynTwins ranks first in reconstruction performance, especially notable in realistic molecules like USPTO molecules and FDA-approved drugs, where other methods struggle to recover most of the original molecules. This suggests that SynTwins is better designed to handle both virtual and real-world molecular structures. Notably, although the approximation process of SynTwins inevitably sacrifices a significant portion of the reconstruction rate, it can generate synthetically accessible analogs that are highly structurally similar to the targets.

In terms of structural similarity, SynTwins also leads across all datasets, producing molecules that are more structurally similar to the target chemicals even when exact reconstruction is not achieved. While ChemProjector performs moderately well, especially with virtual molecules, it falls short in real-world cases. SynFormer, though competitive for the ChEMBL test set, consistently ranks lowest overall. These results highlight the versatility and robustness of SynTwins in generating high-quality molecules across diverse chemical spaces. The ablation study of using different fingerprint size, number of neighbors and depth of the synthesis tree of SynTwins can be found in the **Supporting Information**.

Table 1. The reconstruction rate and the top-k average similarity of generated molecules on the benchmarked datasets.

Test data	Model	Reconstruction rate	Top-k average similarity		
			k = 1	3	5
Virtual molecules	ChemProjector	39.8%	0.8018	0.7554	0.7268
	SynFormer	7%	0.6543	0.6247	0.6070
	SynTwins (this work)	**55%**	**0.8701**	**0.8209**	**0.7992**
ChEMBL molecules	ChemProjector	10.8%	0.5612	0.5256	0.5043
	SynFormer	8.9%	0.5725	0.5397	0.5221
	SynTwins (this work)	**19.6%**	**0.6630**	**0.6222**	**0.6025**
USPTO molecules	ChemProjector	0%	0.4225	0.4103	0.4009
	SynFormer	0%	0.4323	0.4204	0.4114
	SynTwins (this work)	**1.8%**	**0.5298**	**0.5064**	**0.4941**
FDA-approved drugs	ChemProjector	1%	0.4543	0.4382	0.4258
	SynFormer	4%	0.3294	0.3089	0.2979
	SynTwins (this work)	**17%**	**0.6387**	**0.6051**	**0.5873**

Next, we analyze the relationship between the heuristic synthetic accessibility score and the top-5 similarity of molecule analogs generated by SynTwins for the USPTO molecules in **Figure 3**. The synthetic accessibility score of a molecule is calculated by the building block and reaction-aware SAScore (BR-SAScore)[36] using the reactions and building blocks available in this study. For reading clarity, we show the changes in BR-SAScores and molecular similarity of 10 molecules with the lowest BR-SAScores (easy-to-synthesize) and 10 molecules with the highest BR-SAScores (hard-to-synthesize) sampled from the USPTO molecules in **Figure 3a** and **3b**. Similar plots for the FDA-approved drugs, and comparison with ChemProjector and SynFormer for all USPTO and FDA-approved molecules are provided in the **Supporting Information**.
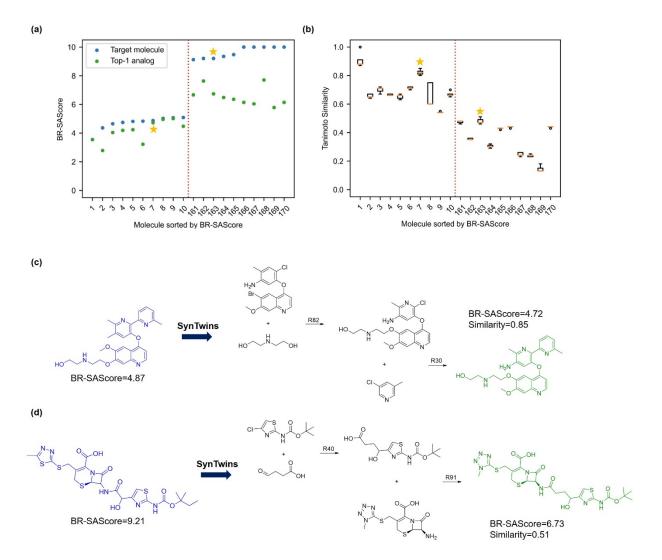
Figure 3. The statistics and examples of synthetically accessible analogs generated by SynTwins from the molecules sampled from the USPTO dataset[32]. The molecules shown as the examples in subfigures c and d are highlighted by star symbols in subfigures a and b. (a) The BR-SAScores of the target molecules and the most similar molecular analogs generated by SynTwins. (b) The box plots of the top-5 similarity of the molecular analogs generated by SynTwins. (c) An example of the most similar molecular analog generated from an easy-to-synthesize target molecule by SynTwins. The target molecule is colored in blue, and the generated molecular analog is colored in green. (d) An example of the most similar molecular analog generated from a hard-to-synthesize target molecule by SynTwins. The target molecule is colored in blue, and the generated molecular analog is colored in green. R82: Williamson reaction, R30: Negishi coupling, R40: Grignard reaction, R91: Amide formation. The full list of the reaction can be found in the **Supporting Information**.

As expected, molecules with lower BR-SAScores are easier for SynTwins to search the synthetically accessible analogs with higher similarity. Notably, the 7 molecules (1 from USPTO-190 molecules and 6 from FDA-approved drugs) having BR-SAScore lower than 4 are all successfully reconstructed by SynTwins. In contrast, SynTwins struggles to generate molecules having high BR-SAScores, resulting molecule analogs with low structural similarities. Furthermore, the generated molecular analogs consistently exhibited lower BR-SAScores than their corresponding target molecules. This decrease was more evident for hard-to-synthesize molecules than for easy-to-synthesize ones.

To visualize the how the Tanimoto similarity aligns with the chemist's view of structural similarity, we show the target molecules and the analogs generated by SynTwins for one easy-to-synthesize and one hard-to-synthesize molecules sampled from the USPTO molecules in **Figure 3c** and **3d**. For easy-to-

synthesize molecules (**Figure 3c**), SynTwins is able to generate a highly structurally similar molecule analog with a 0.85 similarity score. On the other hand, for hard-to-synthesize molecules (**Figure 3d**), SynTwins struggles to generate structurally similar molecules with a 0.51 similarity score. We note that even if the generated molecules are structurally different from the original molecules, especially for the complex molecular structures, their bioactivity can be similar when the crucial functional groups preserved. For instance, the beta-lactam substructure, a well-known functional group for antibiotics, in the target shown in **Figure 3c** is preserved in the generated synthetically accessible analogs despite the low structural similarity
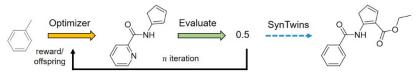
Molecule optimization

Here, we explore the application of embedding SynTwins into the existing molecule optimization methods to generate
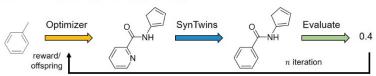
optimized molecules that are synthetically accessible by the available reactions and building blocks. Here, we selected the two leading algorithms for molecule optimization, REINVENT[37] and GraphGA[38], according to the practical molecule optimization (PMO) benchmark[39]. For each algorithm, we designed a synthesizable molecule optimizer variant that converts each optimized molecule into a molecule analog using SynTwins during the optimization process, which guarantees the synthetic accessibility of all the generated molecules (**Figure 4a**). For comparison, we also compare the results by using SynTwins to convert the molecules optimized at the end of the optimization process. The results of the former methods are denoted as "Syn-REINVENT" and "Syn-GraphGA", and the later methods are denoted as REINVENT* and GraphGA* for the variants using REINVENT and GraphGA as molecule optimization backends, respectively. We perform the optimization on 7 similarity-related MPO tasks selected from GuacaMol[40]. We only use the top-1 molecule analog generated from SynTwins with $d = 2$ and $k = 1$

in this experiment. The scoring criteria of each MPO task is given in the **Supporting Information**.
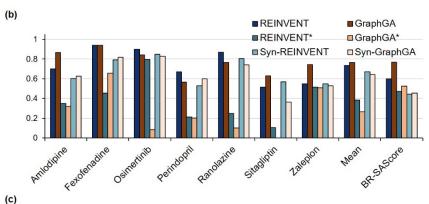
The top-10 scores of the molecules generated by the comparing algorithms for the 7 MPO tasks are shown in **Figure 4b**. Overall, unconstrained REINVENT and GraphGA shows higher top-10 score and higher BR-SAScores than their variants. Nonetheless, these (potentially unsynthesizable) optimized molecules exhibit huge MPO score drops after they were converted to synthetically accessible analogs by SynTwins, with 0.351 and 0.498 difference for REINVENT and GraphGA, respectively. On the other hand, the proposed synthesizable molecule optimizers, Syn-REINVENT and Syn-GraphGA, show slightly lower MPO performance than regular molecule optimizers while guaranteeing their synthetic accessibility and low BR-SAScores. The top-1 and top-100 scores of the 7 MPO tasks are provided in the **Supporting Information**.
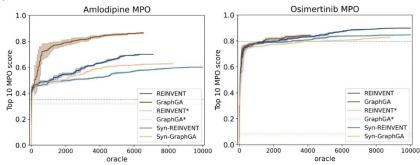
(a)

(b)

(c)

Figure 4. The workflow and results of embedding SynTwins in the molecule optimization tasks. (a) The comparison of the regular molecule optimizer and the synthesizable molecule optimizer. SynTwins can be applied to the final molecule generated by the regular molecule optimizer to convert the optimized molecules into synthetically accessible molecules. (b) The top-10 scores and the BR-SAScores of the compared models on 7 multi-property optimization (MPO) tasks from GuacaMol[40]. The BR-SAScores are rescaled by a factor of 0.1 to match the scale of other metrics. (c) The optimization process of top-10 scores of compared

molecule optimizers on two MPO tasks. The concrete lines represent the top-10 average scores and the shadow near lines are the standard deviations. (c) The relationship between the analog similarity and score difference of the top-100 molecules generated by REINVENT and GraphGA.

We analyze the optimization curves of the optimizers and show two of them in **Figure 4c**. We observed that integrating SynTwins into the optimization algorithms leads to longer converging time than the original algorithms, showing the increased difficulty of optimizing the molecules in a synthetically accessible molecule space. Particularly, we found several molecules generated by the original optimization algorithms collapse to the same molecules during the analog searching process, making the time of optimizing the synthetically accessible molecules significantly longer than that of the regular optimization algorithms. The optimization plots of all the 7 tasks for top-1, top-10 and top-100 molecules can be found in the **Supporting Information**.

Conclusion

We presented SynTwins as a robust and powerful tool for designing synthetically accessible molecule analogs for target molecules. The proposed retrosynthesis-guided molecular analog design framework leverages both retro-reaction and forward-reaction templates to design analogs within a finite set of reactions and building blocks that are commercially available. The key advantages of SynTwins compared to existing methods are twofold. First, unlike previous bottom-up approaches[17,18,25,26,28] that directly sample building blocks based on an arbitrary embedding of the target molecule, SynTwins employs a top-down precursor-searching strategy that aligns with chemists' practice of designing molecular analogs using available building blocks. By mimicking this intuition through a three-step process (retrosynthesis, similar building block searching, and virtual synthesis), SynTwins provides interpretable results and allows for easy optimization to better reflect chemists' needs. Second, SynTwins does not rely on machine learning models[17,18,25,26], making it more robust to variations in training hyperparameters and hardware constraints. For instance, training SynFormer requires over 1,000 GPU hours[26], whereas SynTwins operates efficiently without such computational demands due to an efficient search strategy. Moreover, it can flexibly adapt to different reaction conditions and building block sets without requiring retraining from scratch. SynTwins has the potential to be embedded into molecular optimization workflows, enabling synthesis-aware molecular design. We anticipate that SynTwins will contribute to more efficient molecular design by facilitating the synthesis of viable molecular analogs using readily available reactions and building blocks.
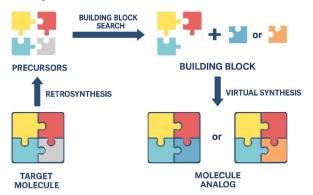
Keywords: synthesizable molecule • molecular analog design • retrosynthesis • reaction template • organic

References

[1]   P. G. Polishchuk, T. I. Madzhidov, A. Varnek, *J Comput Aided Mol Des* **2013**, *27*, 675–679.
[2]   J.-L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722–730.
[3]   B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360–365.
[4]   Y. Du, A. R. Jamasb, J. Guo, T. Fu, C. Harris, Y. Wang, C. Duan, P. Liò, P. Schwaller, T. L. Blundell, *Nat Mach Intell* **2024**, *6*, 589–604.
[5]   W. Gao, C. W. Coley, *J. Chem. Inf. Model.* **2020**, *60*, 5714–5723.
[6]   M. H. S. Segler, M. P. Waller, *Chemistry – A European Journal* **2017**, *23*, 5966–5971.
[7]   C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281–1289.
[8]   S. Chen, Y. Jung, *JACS Au* **2021**, *1*, 1612–1620.
[9]   M. E. Maier, *Org. Biomol. Chem.* **2015**, *13*, 5302–5343.
[10]   A. L. Harvey, *Drug Discovery Today* **2008**, *13*, 894–901.
[11]   R. A. Sheldon, *Green Chem.* **2017**, *19*, 18–43.
[12]   R. A. Sheldon, *ACS Sustainable Chem. Eng.* **2018**, *6*, 32–48.
[13]   M. C. Bryan, P. J. Dunn, D. Entwistle, F. Gallou, S. G. Koenig, J. D. Hayler, M. R. Hickey, S. Hughes, M. E. Kopach, G. Moine, P. Richardson, F. Roschangar, A. Steven, F. J. Weiberth, *Green Chem.* **2018**, *20*, 5082–5103.
[14]   S. K. Gottipati, B. Sattarov, S. Niu, Y. Pathak, H. Wei, S. Liu, S. Liu, S. Blackburn, K. Thomas, C. Coley, J. Tang, S. Chandar, Y. Bengio, in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, **2020**, pp. 3668–3679.
[15]   M. Koziarski, A. Rekesh, D. Shevchuk, A. van der Sloot, P. Gaiński, Y. Bengio, C. Liu, M. Tyers, R. Batey, *Advances in Neural Information Processing Systems* **2024**, *37*, 46908–46955.
[16]   M. Cretu, C. Harris, I. Igashov, A. Schneuing, M. Segler, B. Correia, J. Roy, E. Bengio, P. Liò, **2024**, DOI 10.48550/arXiv.2405.01155.
[17]   J. Noh, D.-W. Jeong, K. Kim, S. Han, M. Lee, H. Lee, Y. Jung, in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, **2022**, pp. 16952–16968.
[18]   W. Gao, R. Mercado, C. W. Coley, **2022**, 25.
[19]   S. Chen, J. Noh, J. Jang, S. Kim, G. H. Gu, Y. Jung, *Acc. Chem. Res.* **2024**, *57*, 1964–1972.
[20]   T. Cover, P. Hart, *IEEE Transactions on Information Theory* **1967**, *13*, 21–27.
[21]   D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
[22]   "Daylight Theory: SMARTS - A Language for Describing Molecular Patterns," can be found under https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, **n.d.**
[23]   C. W. Coley, W. H. Green, K. F. Jensen, *Journal of Chemical Information and Modeling* **2019**, *59*, 2529–2537.
[24]   S. Chen, S. An, R. Babazade, Y. Jung, *Nat Commun* **2024**, *15*, 2250.
[25]   S. Luo, W. Gao, Z. Wu, J. Peng, C. W. Coley, J. Ma, **2024**, DOI 10.48550/arXiv.2406.04628.
[26]   W. Gao, S. Luo, C. W. Coley, **2024**, DOI 10.48550/arXiv.2410.03494.
[27]   M. Hartenfeller, M. Eberle, P. Meier, C. Nieto-Oberhuber, K.-H. Altmann, G. Schneider, E. Jacoby, S. Renner, *J. Chem. Inf. Model.* **2011**, *51*, 3093–3098.
[28]   A. Button, D. Merk, J. A. Hiss, G. Schneider, *Nat Mach Intell* **2019**, *1*, 307–315.
[29]   "Building Blocks - Enamine," can be found under https://enamine.net/building-blocks, **n.d.**
[30]   S. K. Gottipati, B. Sattarov, S. Niu, Y. Pathak, H. Wei, S. Liu, S. Blackburn, K. Thomas, C. Coley, J. Tang, S. Chandar, Y. Bengio, in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, **2020**, pp. 3668–3679.
[31]   A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, A. R. Leach, *Nucleic Acids Research* **2017**, *45*, D945–D954.
[32]   B. Chen, C. Li, H. Dai, L. Song, in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, **2020**, pp. 1608–1616.
[33]   C. for D. E. and Research, "Novel Drug Approvals at FDA," can be found under https://www.fda.gov/drugs/development-approval-process-drugs/novel-drug-approvals-fda, **2025**.
[34]   P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
[35]   D. Bajusz, A. Rácz, K. Héberger, *Journal of Cheminformatics* **2015**, *7*, 20.
[36]   S. Chen, Y. Jung, *Journal of Cheminformatics* **2024**, *16*, 83.
[37]   M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, *Journal of Cheminformatics* **2017**, *9*, 48.
[38]   J. H. Jensen, *Chem. Sci.* **2019**, *10*, 3567–3572.
[39]   W. Gao, T. Fu, J. Sun, C. Coley, *Advances in Neural Information Processing Systems* **2022**, *35*, 21342–21357.
[40]   N. Brown, M. Fiscato, M. H. S. Segler, A. C. Vaucher, *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108.

Entry for the Table of Contents

SynTwins enables the design of synthesizable molecular analogs by combining retrosynthesis with precursors replacement. The figure illustrates a target molecule decomposed into precursors, including grey (commercially unavailable) fragments. SynTwins identifies similar colored (commercially available) building blocks to reconstruct the molecule using chemical reactions, generating analogs that preserve core structure while ensuring synthetic accessibility.