Electronic Supplementary Information (ESI):

SynTwins: A Retrosynthesis-Guided Framework for Synthesizable Molecular Analog Generation

Shuan Chen,[ab] Gunwook Nam,[a] Alan Aspuru-Guzik,[c] and Yousung Jung*[abd]

a. Department of Chemical and Biological Engineering, and Institute of Chemical Processes, Seoul National University, 1 Gwanak-ro, Seoul, South Korea

b. Institute of Engineering Research, Seoul National University, 1 Gwanak-ro, Seoul, South Korea

c. Department of Chemistry, Department of Computer Science, University of Toronto, Vector Institute for Artificial Intelligence, Canadian Institute for Advanced Research, Toronto, Ontario, Canada

d. Interdisciplinary Program in Artificial Intelligence, Seoul National University, 1 Gwanak-ro, Seoul, South Korea

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

Table of Contents

S1. Method and Materials

S1.1. Retro-reaction templates curation

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 chlorine (Cl), bromine (Br) or iodine (I) to run the reaction, it makes direct derivation of retro-reaction templates by simply reversing reactants and products infeasible (**Fig. S1**).

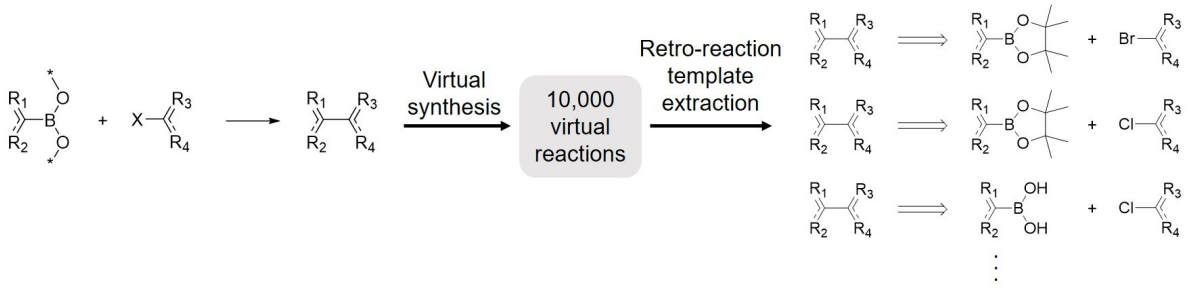


Fig. S1. The process of deriving retro-reaction templates from the available forward reaction templates and building blocks.

In this work, we collected a total of 122 forward-reaction templates, including 58 from Hartenfeller et al.[1] and 64 from Button et al.[2]. Because the style of reaction naming and reaction SMIRKS crafting is different between these two literatures, it is nontrivial to remove the duplicated reactions only by seeing the reaction name or reaction SMRIKS. Therefore, we remove the duplicated reactions by comparing the retro-reaction templates generated from each forward-reaction template, since the style of retro-reaction template is unified by the same reaction template extraction function.

If one reaction template T_i results to a set of retro-reaction templates $\{T_{r,i}\}$ and another reaction template T_j results to a set of retro-reaction templates $\{T_{r,j}\}$, we define these two reaction templates are duplicated if a set retro-reaction templates derived from one reaction template is a subset of the retro-reaction templates derived from another reaction template. In other words, T_i and T_j is duplicated if $\{T_{r,j}\} \subseteq \{T_{r,i}\}$ or vice versa. In this case, we removed the forward-reaction template resulting to fewer retro-reaction templates and thus resulted to 101 reaction templates. After selecting the retro-reaction templates derived from these reaction templates and remove the ones with low frequency, 1,163 retro-reaction templates were left in this work to accelerate the run time of SynTwins. Note that the retro-reaction templates are only used in SynTwins, not for other comparing baselines.

Note that the retro-reaction template extraction algorithm is based on RDChiral[3], which was designed to handle the chirality of the chemical reactions. Nonetheless, since none of the forward-reaction templates collected in this work includes stereoselectivity information, none of the retro-reaction templates, extracted from the reactions virtually synthesized using these forward-reaction templates, used in this work includes stereoselectivity information either.

S1.2. Single-step retrosynthesis

We use RDKit python package[4] to convert the SMRIKS-based retro-reaction template to RDKit reaction objects. The RDKit reaction object recognizes the matched scaffold in the target molecule and converts the matched product scaffold to the reactant scaffolds. Therefore, a set of reactants can be obtained by applying all the compatible RDKit reaction objects to the target molecule.

Because the number of compatible retro-reactions may be a large (over 100 for big molecules), we designed a computational approximation to reduce the number of generated precursors. We notice that for the same forward-reaction, the reactants generated from different retro-reaction templates would collapse to the same available building blocks in many cases. For example, the Suzuki Coupling reaction shown in **Fig. S1** will generate precursors with bromide and chloride at the halide side. Nonetheless, if only the bromide-based building block is available, the chloride-based precursors will eventually be converted to the bromide-based building block. Thus, we only keep the precursors from k sampled retro-reactions generated from each forward-reaction template to accelerate the computation of SynTwins.

S1.3. K-nearest neighbor (kNN)

We apply k-nearest neighbor algorithm available in scikit-learn[5] python package to search for the structurally similar building blocks from the generated precursors after synthesis planning. We use Extended Connectivity Fingerprints[6] with radius 2 (ECFP4) to embed each molecule into a 1024-dimensional feature and building a ball tree using the building block embeddings for each forward-reaction template. By limiting the kNN search within the ball tree of the reaction template, the building blocks $\{R_{twin}\}$ searched by the kNN algorithm are guaranteed to carry the necessary functional groups in the generated precursor R_{ref} and thus can be successfully used in the later virtual synthesis.

S1.4. Test molecule curation

To evaluate the capability of generating synthetically accessible analogs of the proposed and baseline methods, we curated 1,000 virtually synthesized molecules and collect 1,000 molecules from ChEMBL database, 170 molecules from USPTO-190 dataset[7], and 100 molecules from FDA-approved drugs[8] from 2021 to 2024. For the virtually synthesized molecules, we synthesized the molecules by randomly sampling the available building blocks and reactions up to 3 to 5 steps. For ChEMBL molecules, we use the same 1,000 molecule set provided by Gao et al[9]. By filtering out the molecules having over 50 heavy atoms among the USPTO-190 molecules and FDA-approved drugs, we obtained 170 and 100 molecules from the two test sets, respectively.

S1.5. Training ChemProjector and SynFormer

To compare the SynTwins with existing models, we trained ChemProjector[10] and SynFormer[11] using the 101 reaction templates and 150,560 building blocks curated in this work according to their GitHub instructions. We trained each model using 4 A100 GPUs for 48 hours. The training loss and validation loss of both models converge by the end of the training process.

S2. The detailed computational process of SynTwins

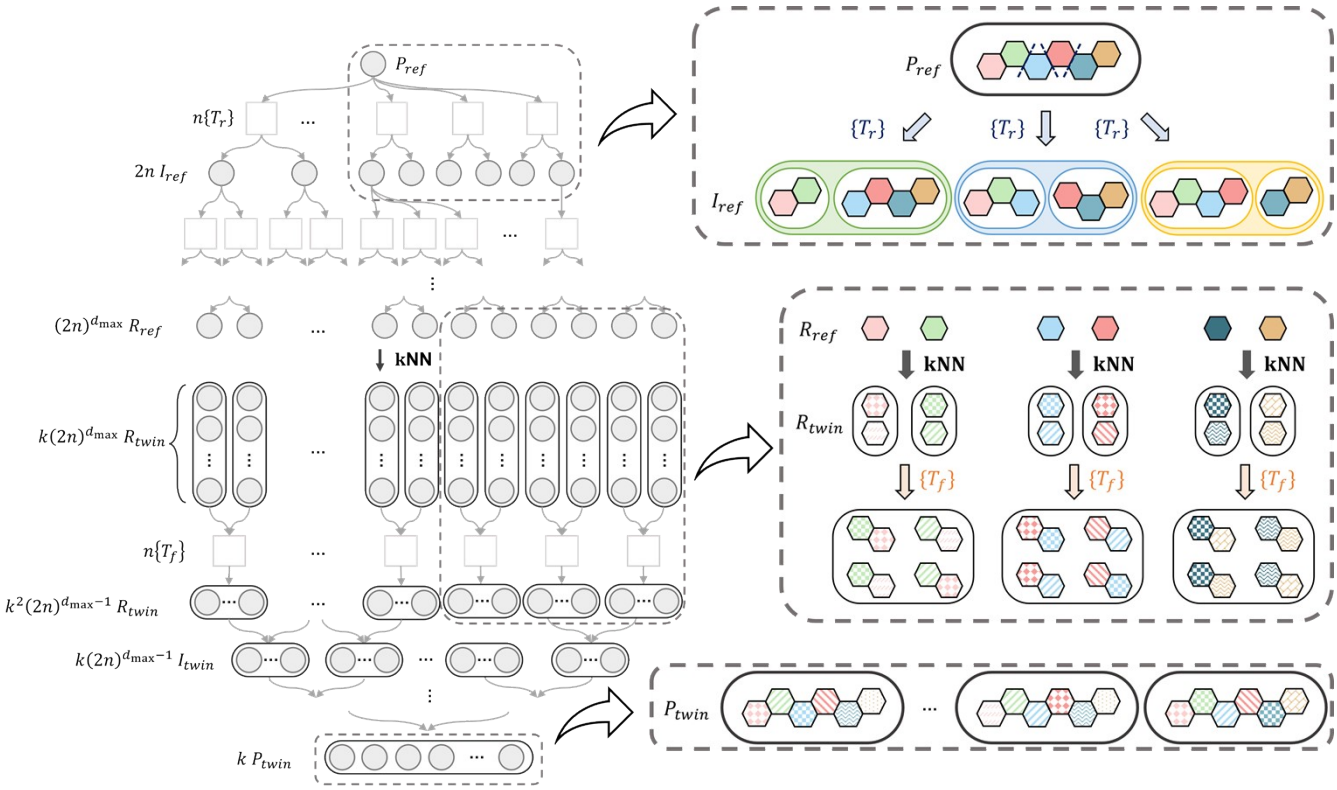


Fig. S2. The synthesis tree and estimated number of molecules generated at each stage of the *SynTwins*. The target molecule and its analogs are denoted as P_{ref} and P_{twin} respectively. Synthesis intermediates are labeled as I_{ref} with their structurally similar counterparts denoted as I_{twin}. Precursors of the target molecules are referred to as R_{ref}, and the building blocks found by the kNN algorithm are denoted as R_{twin}. Detailed zoom-in examples are illustrated in the dashed box.

S3. The chemical reactions used in this work

Table S1. The chemical reactions used in this work. The names of the reactions are collected from the original papers, Hartenfeller et al.[1] and Button et al.[2].

Reaction number	Reaction name	Reference
R0	Pictet-Spengler	[1]
R1	benzimidazole_derivatives_carboxylic-acid/ester	[1]
R2	benzimidazole_derivatives_aldehyde	[1]
R3	benzothiazole	[1]
R4	benzoxazole_arom-aldehyde	[1]
R5	benzoxazole_carboxylic-acid	[1]
R6	thiazole	[1]
R7	Niementowski_quinazoline	[1]
R8	tetrazole_terminal	[1]
R9	tetrazole_connect_regioisomere_2	[1]
R10	Huisgen_Cu-catalyzed_1,4-subst	[1]
R11	Huisgen_disubst-alkyne	[1]
R12	1,2,4-triazole_acetohydrazide	[1]
R13	1,2,4-triazole_carboxylic-acid/ester	[1]
R14	3-nitrile-pyridine	[1]
R15	spiro-chromanone	[1]
R16	pyrazole	[1]
R17	phthalazinone	[1]
R18	Paal-Knorr pyrrole	[1]
R19	triaryl-imidazole	[1]
R20	Fischer indole	[1]
R21	Friedlaender chinoline	[1]
R22	benzofuran	[1]
R23	benzothiophene	[1]
R24	indole	[1]
R25	oxadiazole	[1]
R26	Williamson ether	[1]
R27	reductive amination	[1]
R28	Suzuki	[1]
R29	piperidine_indole	[1]
R30	Negishi	[1]
R31	Mitsunobu_imide	[1]
R32	Mitsunobu_phenole	[1]
R33	Mitsunobu_sulfonamide	[1]
R34	Mitsunobu_tetrazole_1	[1]
R35	Mitsunobu_tetrazole_4	[1]
R36	Heck_terminal_vinyl	[1]

R37	Heck_non-terminal_vinyl	1
R38	Stille	1
R39	Grignard_carbonyl	1
R40	Grignard_alcohol	1
R41	Sonogashira	1
R42	sulfon_amide	1
R43	N-arylation_heterocycles	1
R44	Wittig	1
R45	Buchwald-Hartwig	1
R46	imidazole	1
R47	decarboxylative_coupling	1
R48	heteroaromatic_nuc_sub	1
R49	thiourea	1
R50	Bischler-Napieralski	2
R51	Pictet-Gams	2
R52	Pictet-Spengler-6-membered-ring	2
R53	Pictet-Spengler-5-membered-ring	2
R54	Bischler-Indole	2
R55	Benzimidazol_formation	2
R56	Aminothiazol_formation	2
R57	Benzoxazol_formation	2
R58	Benzothiazol_formation	2
R59	Rap-Stoermer	2
R60	Niementowski	2
R61	Quinazolinone_formation	2
R62	Tetrahydro-Indole_formation	2
R63	3-nitrile_pyridine	2
R64	Huisgen_1-3_Dipolar_Cycloaddition	2
R65	Huisgen_1_3_Dipolar_Cycloaddition_double_bond	2
R66	Diels-Alder	2
R67	Diels-Alder-Alkyne	2
R68	Spiro-piperidine_formation	2
R69	Pyrazol_formation	2
R70	Phthalazinone	2
R71	Fischer_indole	2
R72	Friedlaender_chinoline_formation	2
R73	Peachmann_coumarine	2
R74	Benzofuran_formation	2
R75	Imidazol-Acetamid	2
R76	Dieckmann_5-ring	2
R77	Dieckmann_6-ring	2

R78	Flavone_formation	[2]
R79	Oxadiazole_formation	[2]
R80	Michael_addition	[2]
R81	Cross_Claissen	[2]
R82	Williamson_ether	[2]
R83	Ester_formation	[2]
R84	Reductive_amination-Ketone	[2]
R85	Suzuki_coupling	[2]
R86	Piperidine_and_Indole	[2]
R87	Negishi	[2]
R88	Mitsunobu_imide	[2]
R89	Mitsunobu_sulfonic_amide	[2]
R90	Heck	[2]
R91	Amide_formation	[2]
R92	Ketone_formation	[2]
R93	Ar-Imidazole_formation	[2]
R94	Alkyne_alkylation	[2]
R95	Alkyne_acylation	[2]
R96	FGI_Acyl_chloride	[2]
R97	FGI_bromination	[2]
R98	FGI_sulfonyl_chloride	[2]
R99	FGA_alpha_bromination	[2]
R100	FGI_Rosenmund-von_Braun	[2]

S4. Ablation study of SynTwins

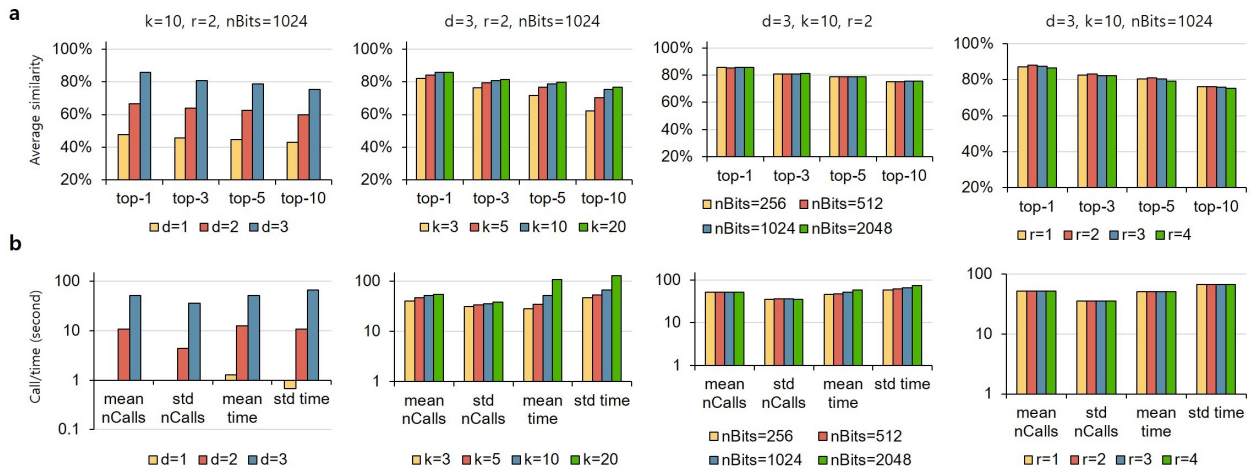


Fig. S3. The ablation study of SynTwins using different depth d, number of neighbors in kNN k, size of the fingerprints vector $nBits$, and fingerprints radius r on the virtually synthesized molecules. (a) top-k average similarities. (b) The number of retrosynthesis call and computational time (second per molecule) for each target molecule.

S5. Results of USPTO molecules and FDA-approved drugs

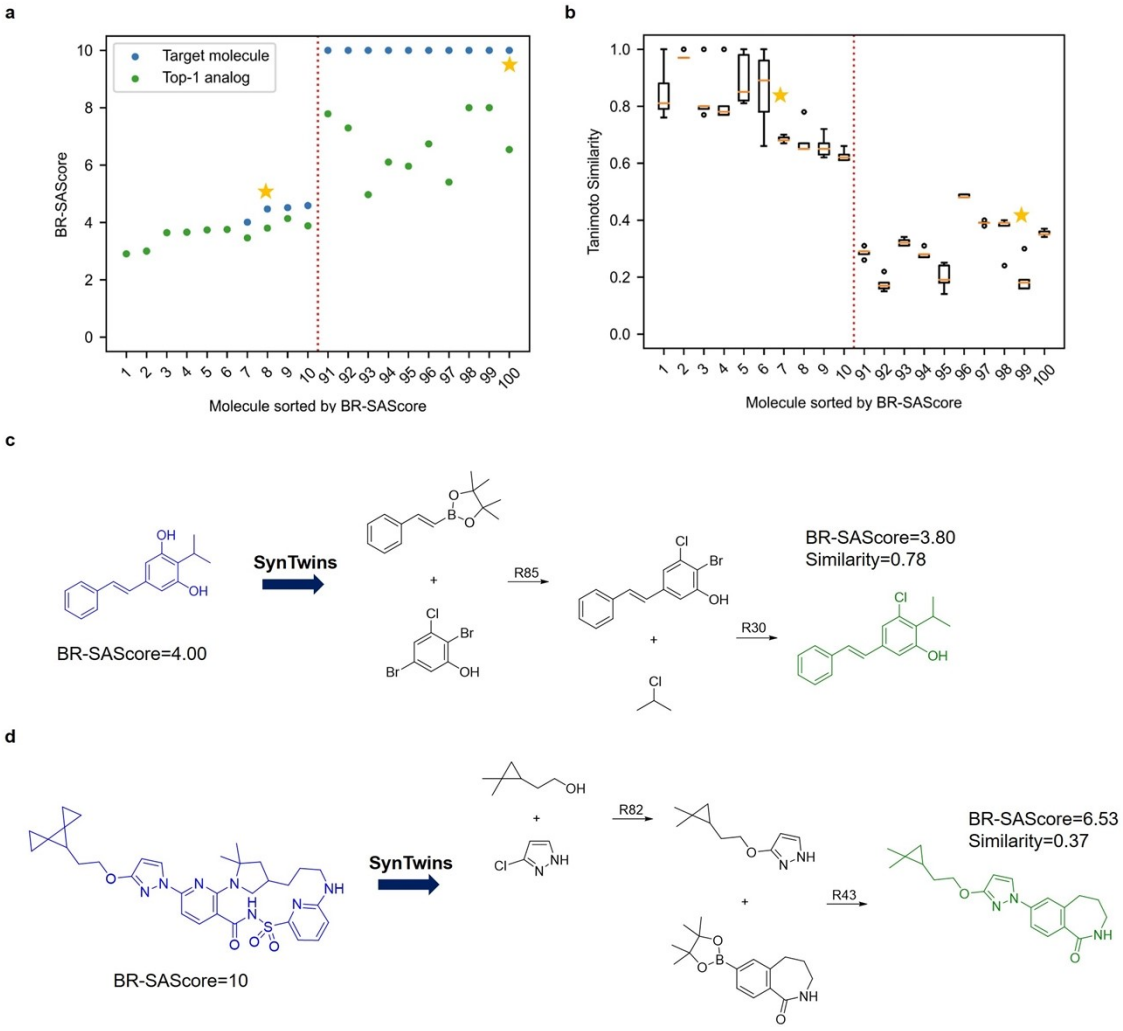


Fig. S4. The statistics and examples of synthetically accessible analogs generated by SynTwins from the molecules sampled from the FDA-approved drugs[8]. The molecules shown as the examples in subfigures c and d are highlighted by star symbols in subfigures a and b. (a) The BR-SAScores of the target molecules and the most similar molecular analogs generated by SynTwins. (b) The box plots of the top-10 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. R85: Suzuki coupling, R30: Negishi coupling, R82: Williamson reaction, R43: N-arylation of heterocycles.

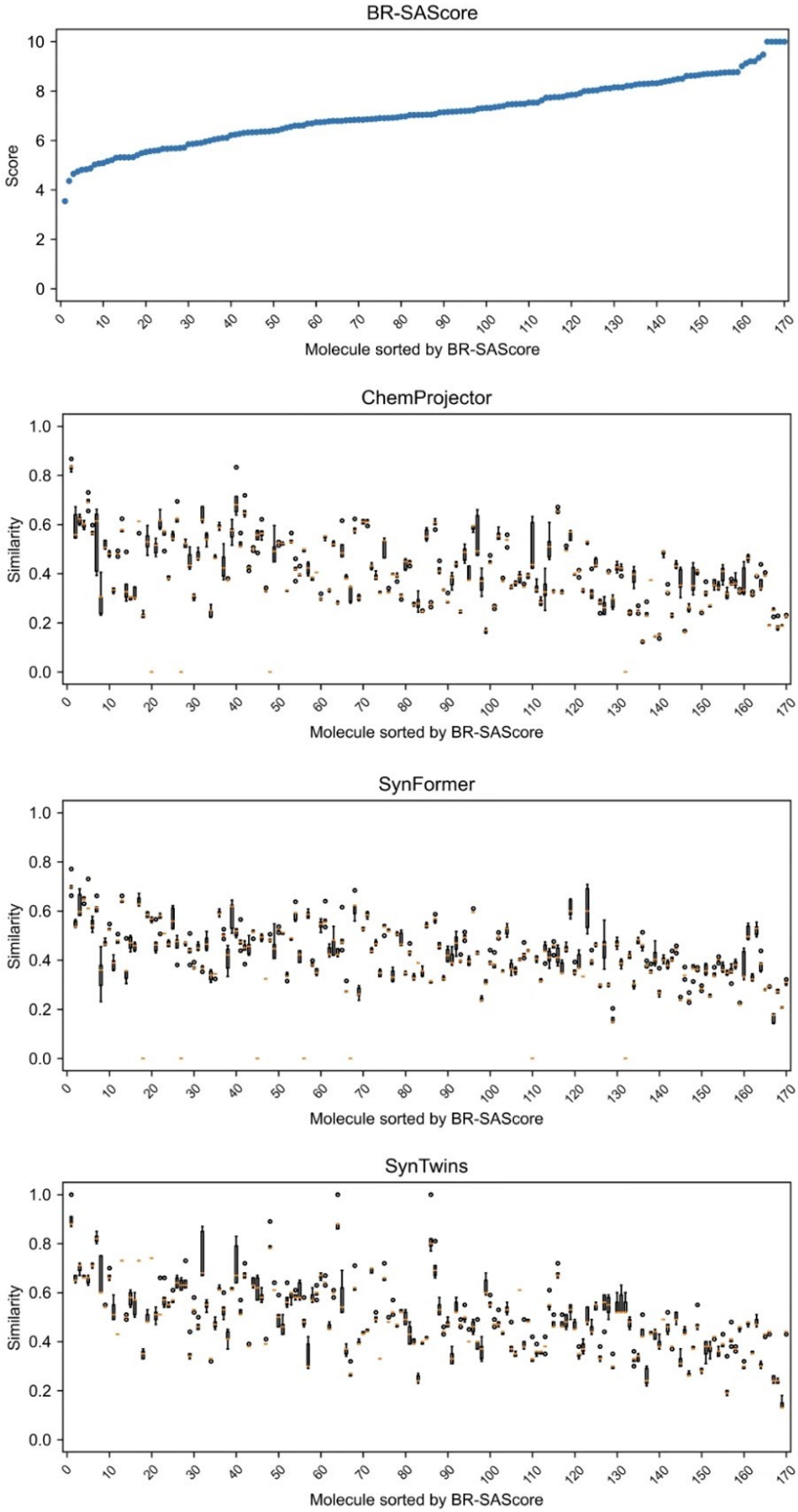


Fig. S5. The BR-SAScores of the USPTO molecules and the structural similarity between top-5 molecule analogs generated by ChemProjector[10], SynFormer[11] and SynTwins (this work).

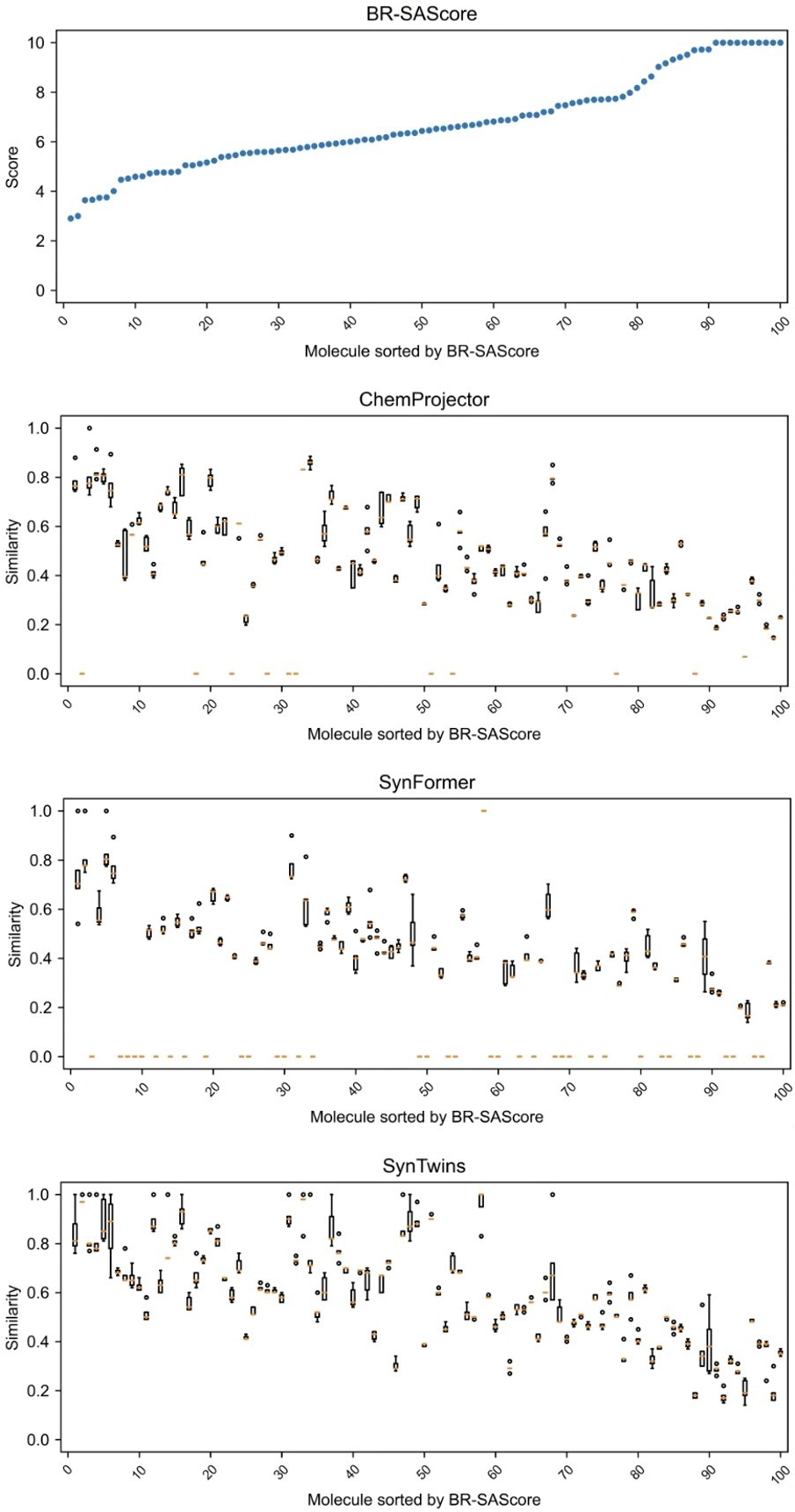


Fig. S6. The BR-SAScores of the FDA-approved drugs and the structural similarity between top-5 molecule analogs generated by ChemProjector[10], SynFormer[11] and SynTwins (this work).

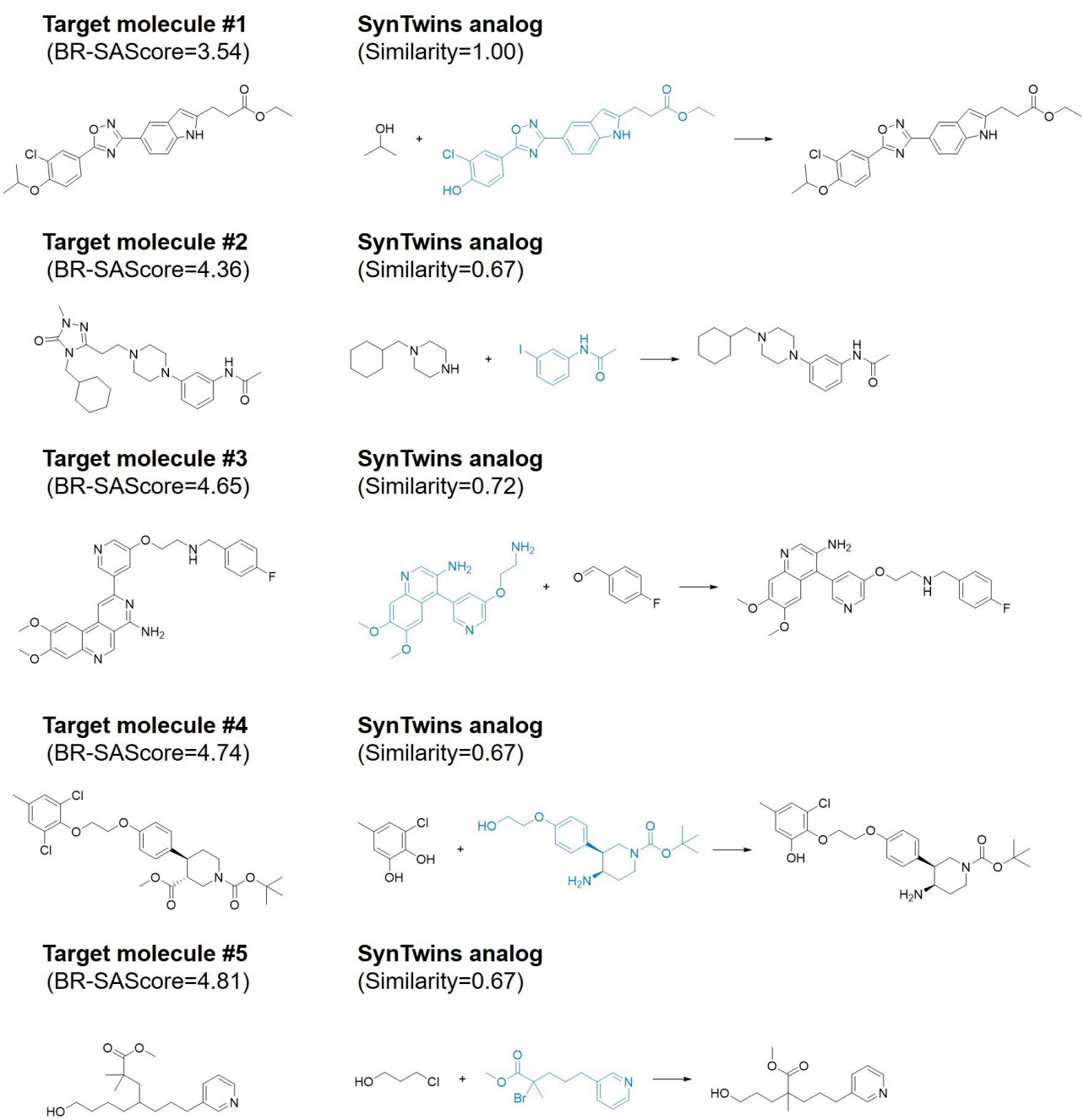


Fig. S7. The target molecules and the SynTwins analogs sampled from USPTO molecules with top-1 to top-5 lowest BR-SAScore. The molecules that require additional synthesis are highlighted in blue color.

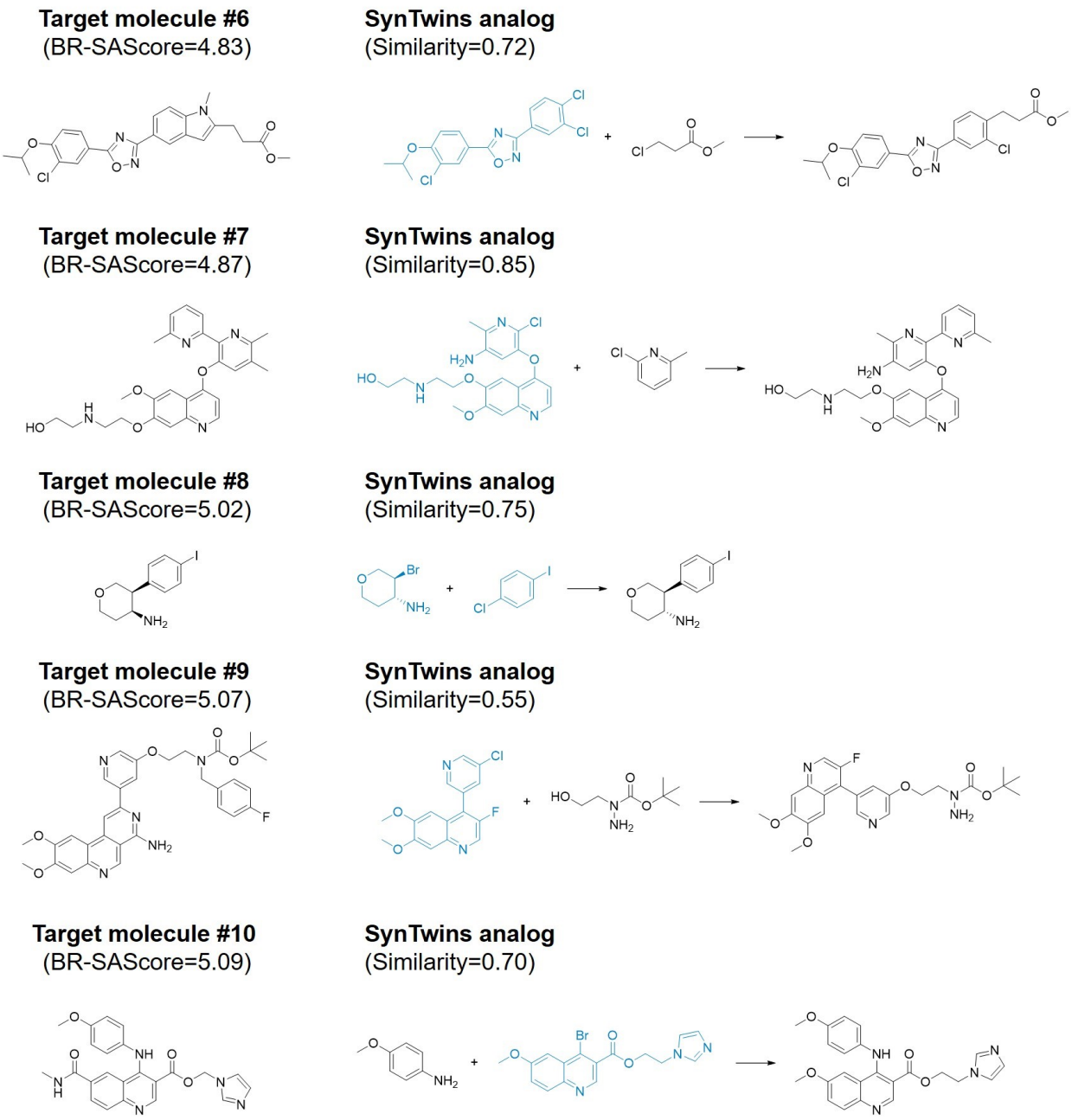
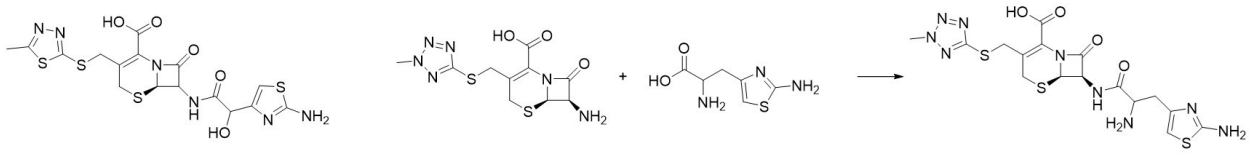


Fig. S8. The target molecules and the SynTwins analogs sampled from USPTO molecules with top-6 to top-10 lowest BR-SAScore. The molecules that require additional synthesis are highlighted in blue color.

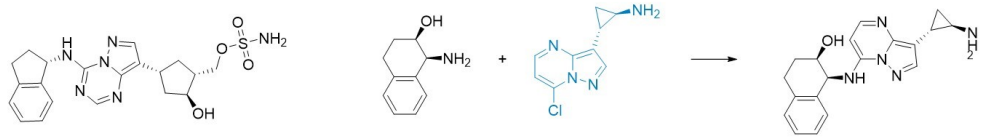
Target molecule #91
(BR-SAScore=9.12)

SynTwins analog
(Similarity=0.48)



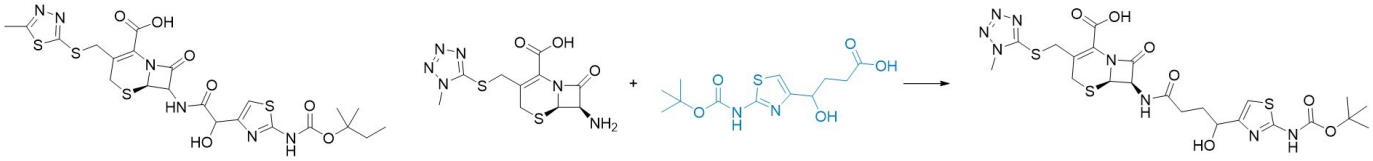
Target molecule #92
(BR-SAScore=9.20)

SynTwins analog
(Similarity=0.36)



Target molecule #93
(BR-SAScore=9.21)

SynTwins analog
(Similarity=0.51)



Target molecule #94
(BR-SAScore=9.34)

SynTwins analog
(Similarity=0.32)

Target molecule #95
(BR-SAScore=9.47)

SynTwins analog
(Similarity=0.43)

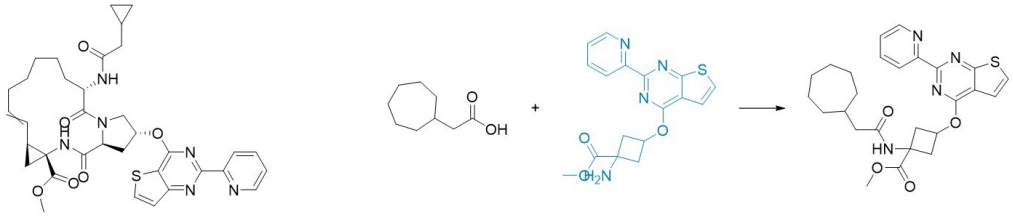


Fig. S9. The target molecules and the SynTwins analogs sampled from USPTO molecules with top-1 to top-5 highest BR-SAScore. The molecules that require additional synthesis are highlighted in blue color.

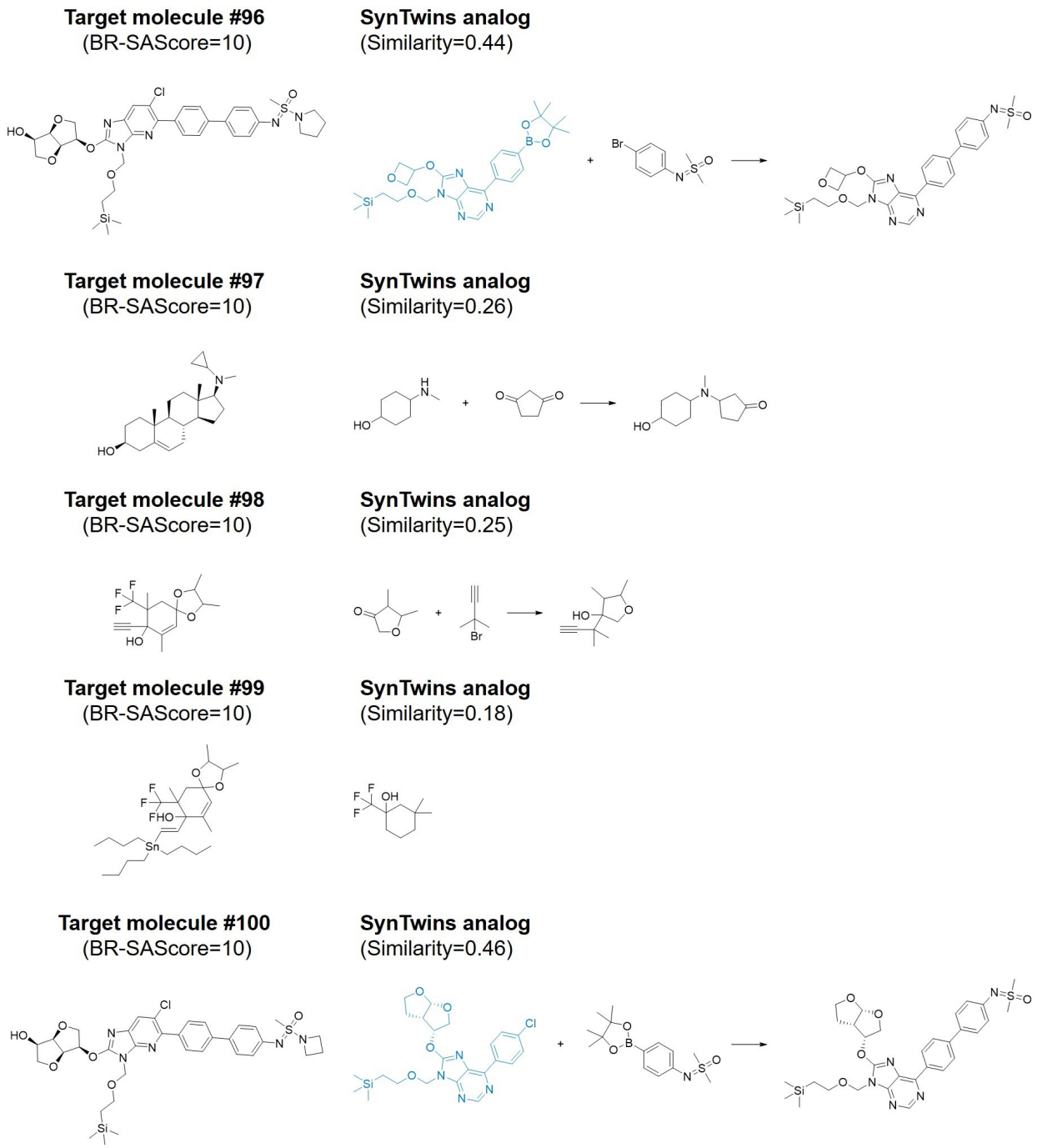


Fig. S10. The target molecules and the SynTwins analogs sampled from USPTO molecules with top-6 to top-10 highest BR-SAScore. The molecules that require additional synthesis are highlighted in blue color.

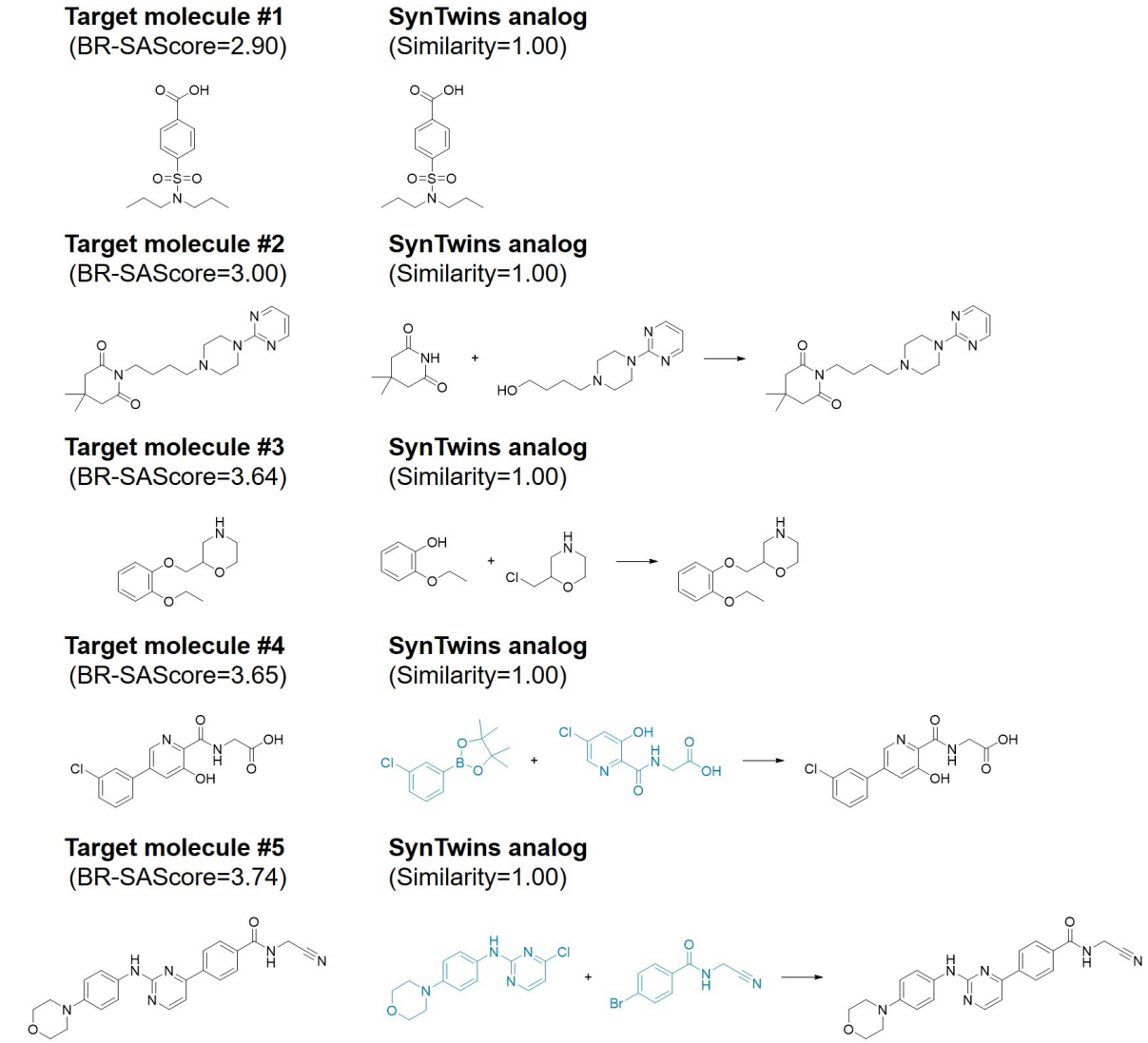


Fig. S11. The target molecules and the SynTwins analogs sampled from FDA-approved drugs with top-1 to top-5 lowest BR-SAScore. The molecules that require additional synthesis are highlighted in blue color.

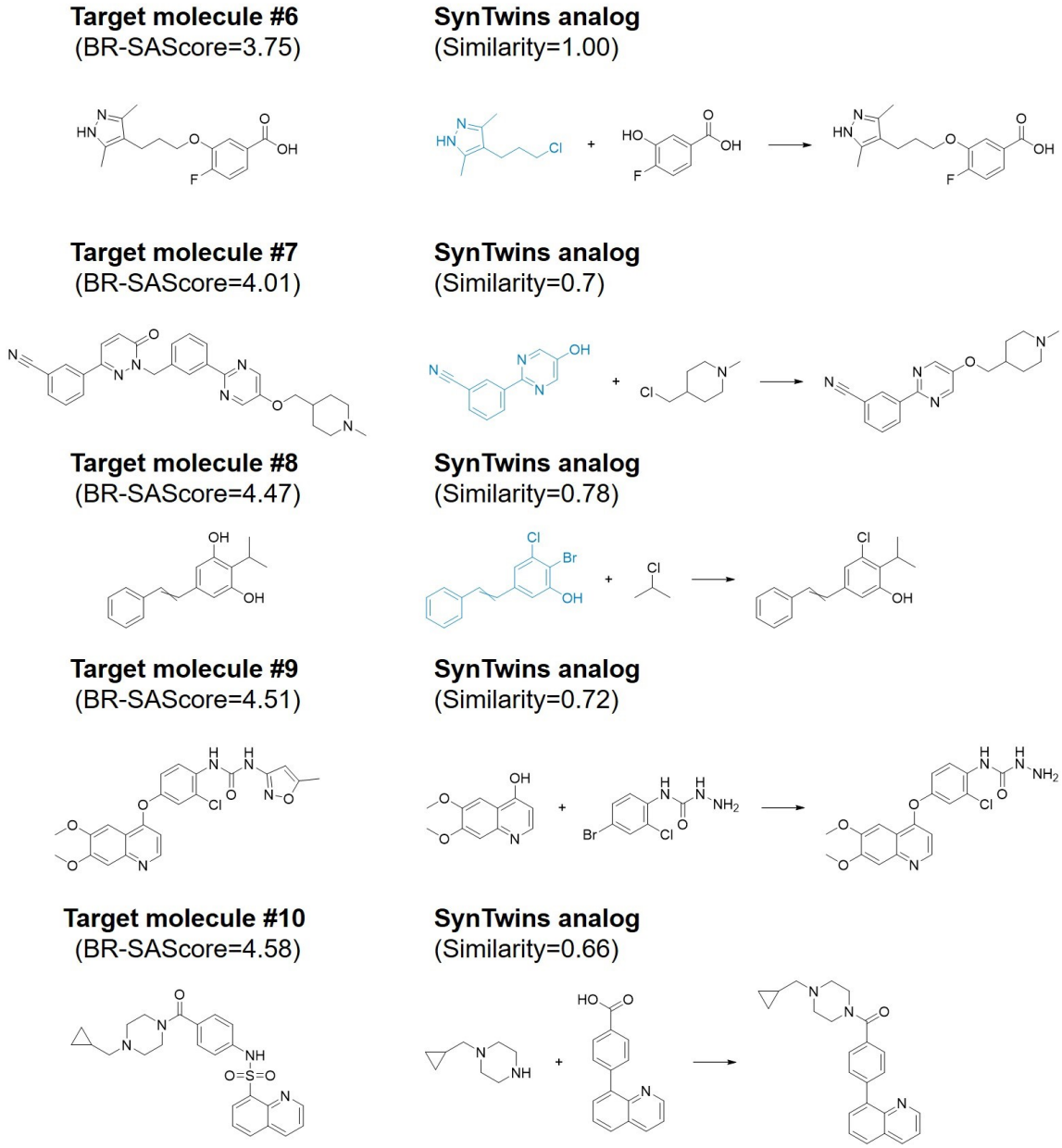
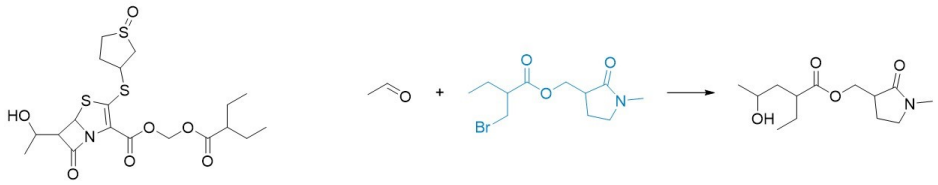


Fig. S12. The target molecules and the SynTwins analogs sampled from FDA-approved drugs with top-6 to top-10 lowest BR-SAScore. The molecules that require additional synthesis are highlighted in blue color.

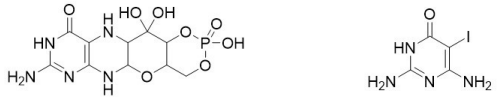
Target molecule #91
(BR-SAScore=10)

SynTwins analog
(Similarity=0.31)



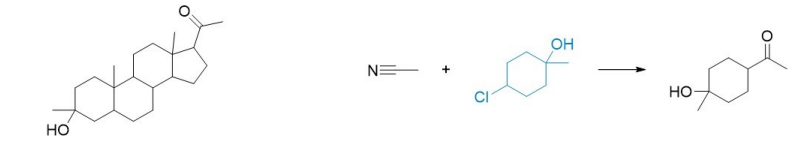
Target molecule #92
(BR-SAScore=10)

SynTwins analog
(Similarity=0.22)



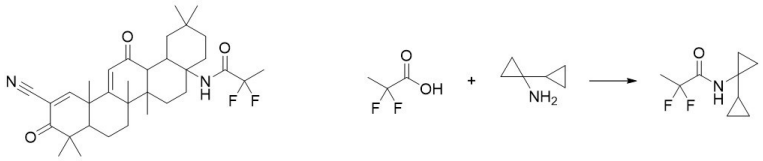
Target molecule #93
(BR-SAScore=10)

SynTwins analog
(Similarity=0.34)



Target molecule #94
(BR-SAScore=10)

SynTwins analog
(Similarity=0.31)



Target molecule #95
(BR-SAScore=10)

SynTwins analog
(Similarity=0.25)

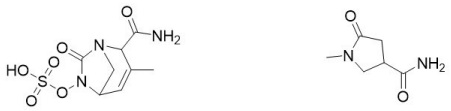


Fig. S13. The target molecules and the SynTwins analogs sampled from FDA-approved drugs with top-6 to top-10 highest BR-SAScore. The molecules that require additional synthesis are highlighted in blue color.

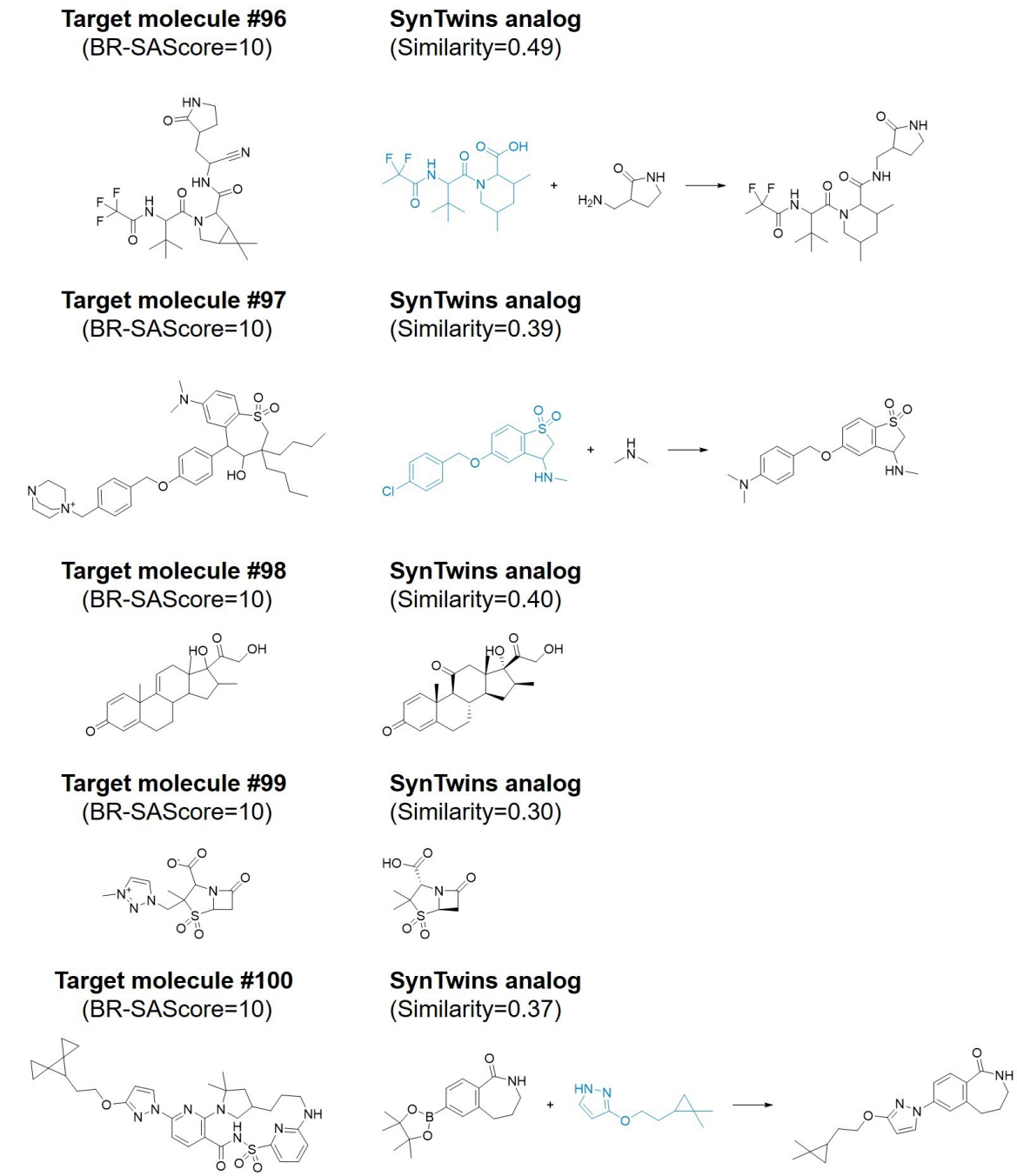


Fig. S14. The target molecules and the SynTwins analogs sampled from FDA-approved drugs with top-1 to top-5 highest BR-SAScore. The molecules that require additional synthesis are highlighted in blue color.

S6. Failure-mode analysis for USPTO molecules

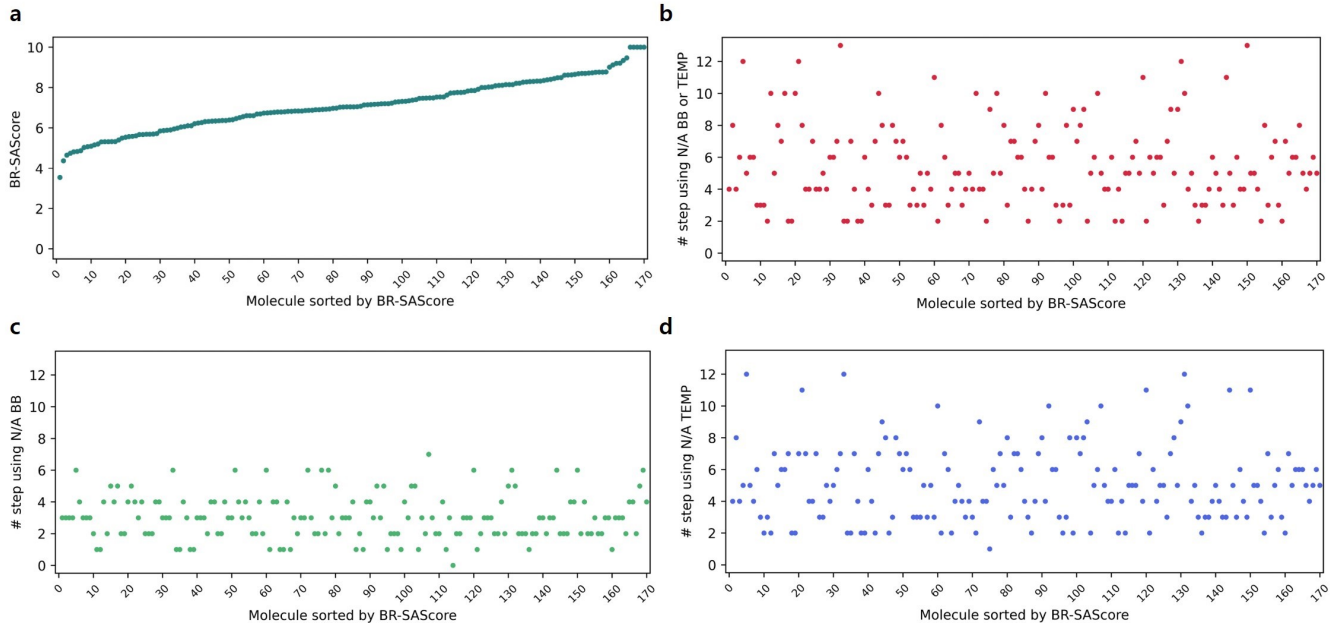


Fig. S15. The failure-mode analysis for the 170 USPTO molecules. (a) The BR-SAScore for each target molecule. (b) The number of reaction steps using either unavailable building blocks (N/A BB) or reaction templates (N/A TEMP) in this study. (c) The number of reaction steps using unavailable building blocks in this study. (d) The number of reaction steps using unavailable reaction templates in this study.

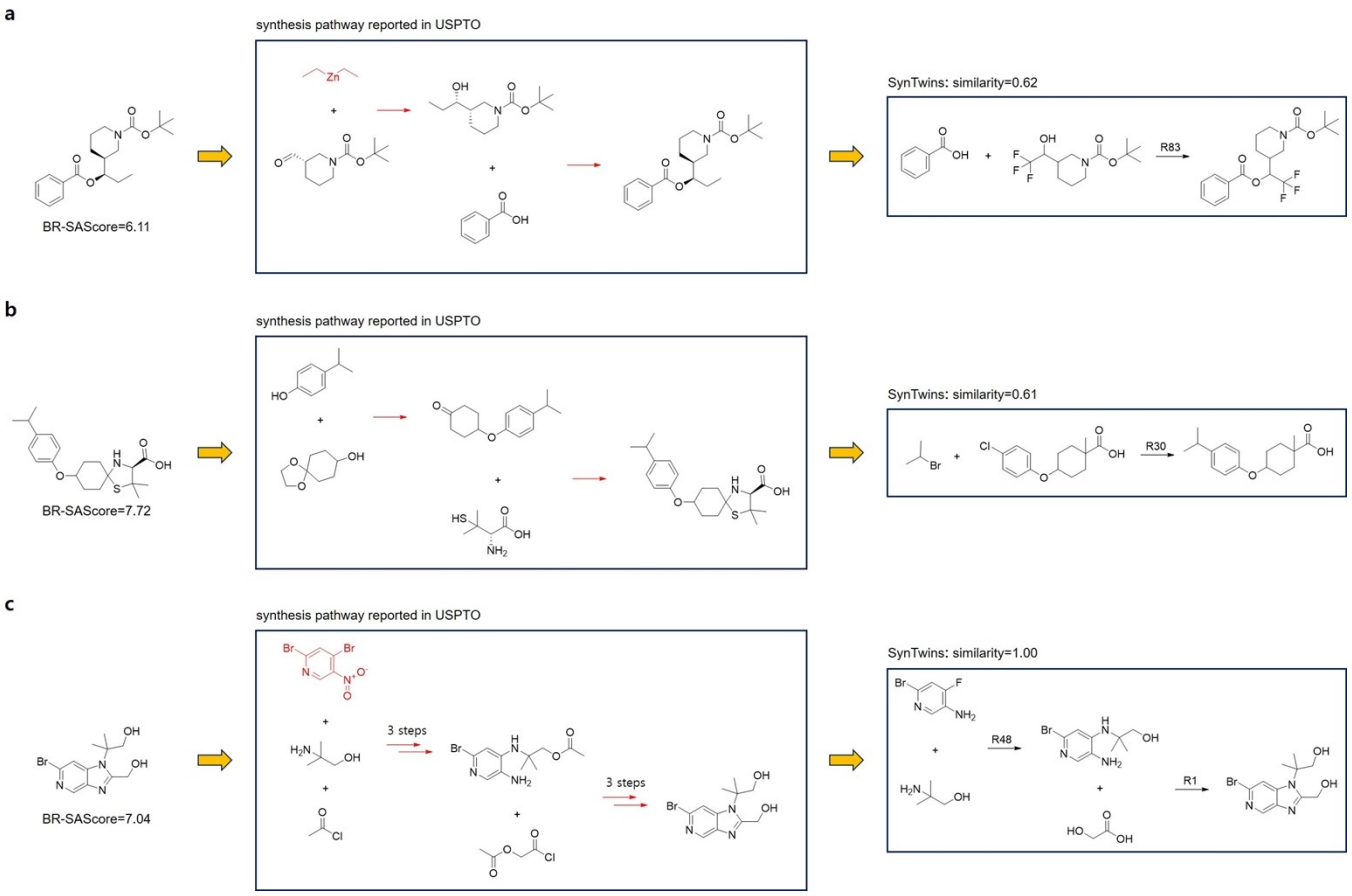


Fig. S16. The examples of USPTO molecules and their most similar analogs generated by SynTwins. (a) A molecule with a reported synthesis pathway using unavailable building blocks and reactions. (b) A molecule with a reported synthesis pathway using available building blocks and unavailable reactions. (c) A molecule with a reported synthesis pathway using unavailable building blocks and reactions, but successfully reconstructed by SynTwins using a different synthesis pathway. The unavailable building blocks and reactions are highlighted in red.

S7. Scoring functions of multi-property optimization (MPO)

Table S2. The scoring functions of each MPO task as described in GuacaMol[12]. The structural similarity (sim), Topological polar surface area (TPSA), and partition coefficient (logP) are calculated by RDKit[13].

Task name	Scoring functions
Amlodipine MPO	sim(amlodipine, ECFP4), number rings
Fexofenadine MPO	sim(fexofenadine, AP), TPSA, logP
Osimertinib MPO	sim(osimertinib, ECFP4), sim(osimertinib, ECFP6), TPSA, logP
Perindopril MPO	sim(perindopril, ECFP4), number aromatic rings
Ranolazine MPO	sim(ranolazine, AP), TPSA, logP, number of fluorine atoms
Sitagliptin MPO	sim(sitagliptin, ECFP4), TPSA, logP, isomer($C_{16}H_{15}F_6N_5O$)
Zaleplon MPO	sim(zaleplon, ECFP4), isomer($C_{19}H_{17}N_3O_2$)

S8. More results of Multi-property optimization (MPO)

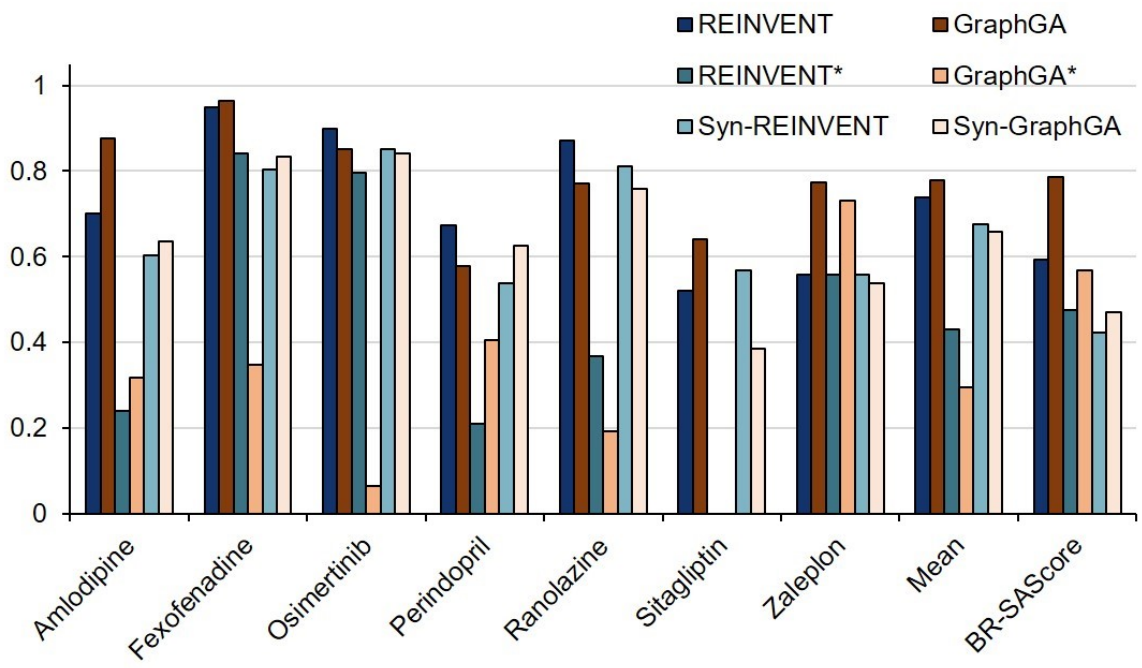


Fig. S17. The MPO results of top-1 molecules generated by REINVENT, GraphGA, and their variants. The BR-SAScores are rescaled by a factor of 0.1 to match the scale of other metrics.

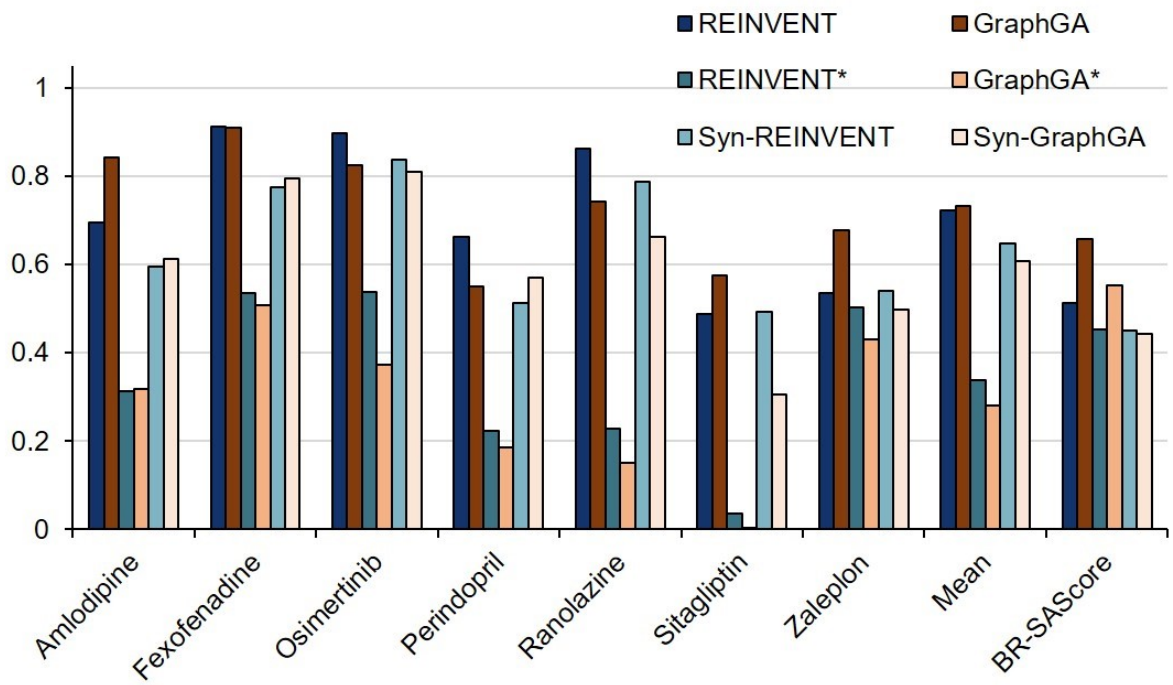


Fig. S18. The MPO results of top-100 molecules generated by REINVENT, GraphGA, and their variants. The BR-SAScores are rescaled by a factor of 0.1 to match the scale of other metrics.

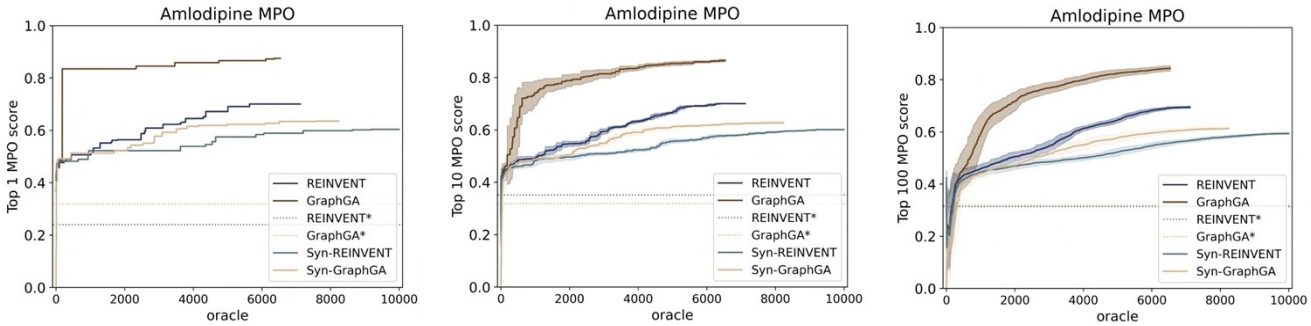


Fig. S19. The top-1 scores (right), top-10 scores (middle), and top-100 scores during the optimization process of REINVENT, GraphGA, Syn-REINVENT, Syn-GraphGA on Amlodipine MPO.

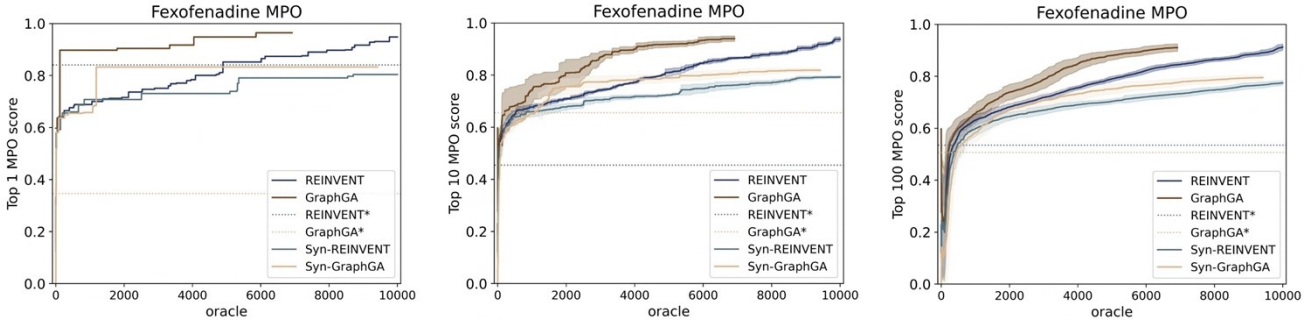


Fig. S20. The top-1 scores (right), top-10 scores (middle), and top-100 scores during the optimization process of REINVENT, GraphGA, Syn-REINVENT, Syn-GraphGA on Fexofenadine MPO.

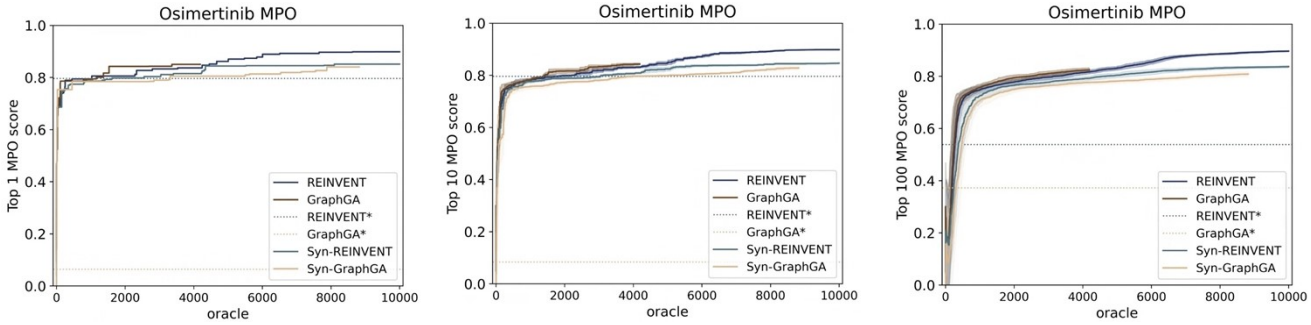


Fig. S21. The top-1 scores (right), top-10 scores (middle), and top-100 scores during the optimization process of REINVENT, GraphGA, Syn-REINVENT, Syn-GraphGA on Osimertinib MPO.

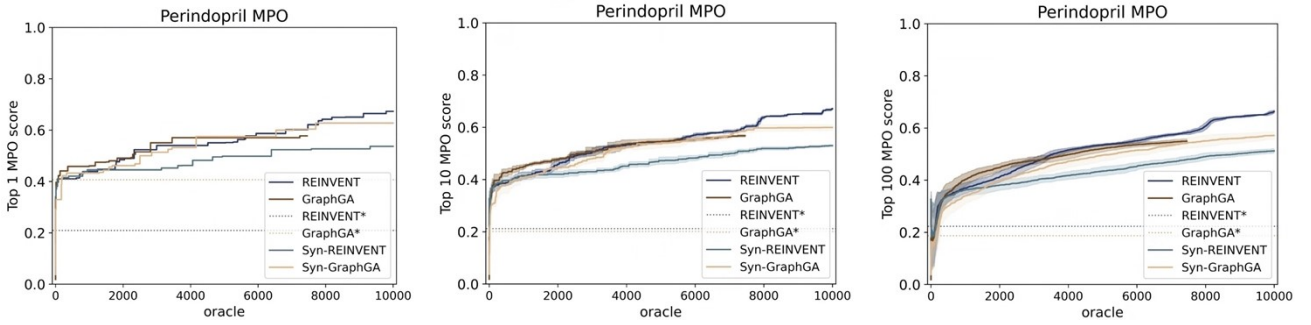


Fig. S22. The top-1 scores (right), top-10 scores (middle), and top-100 scores during the optimization process of REINVENT, GraphGA, Syn-REINVENT, Syn-GraphGA on Perindopril MPO.

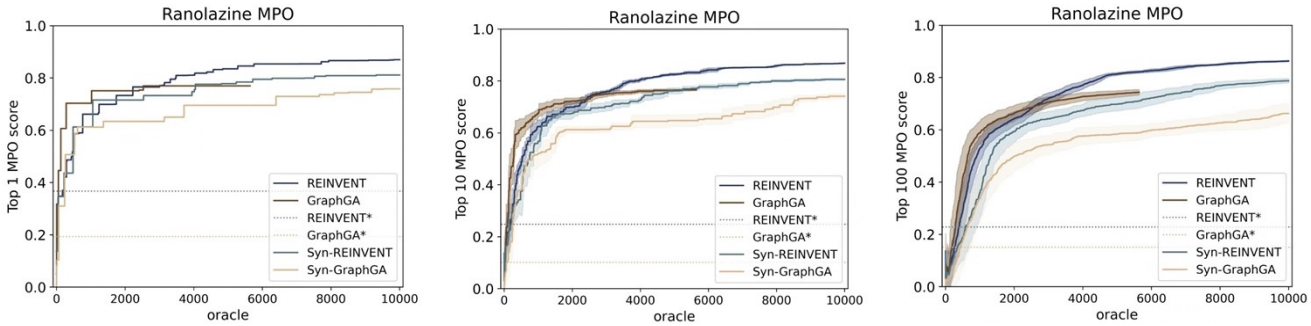


Fig. S23. The top-1 scores (right), top-10 scores (middle), and top-100 scores during the optimization process of REINVENT, GraphGA, Syn-REINVENT, Syn-GraphGA on Ranolazine MPO.

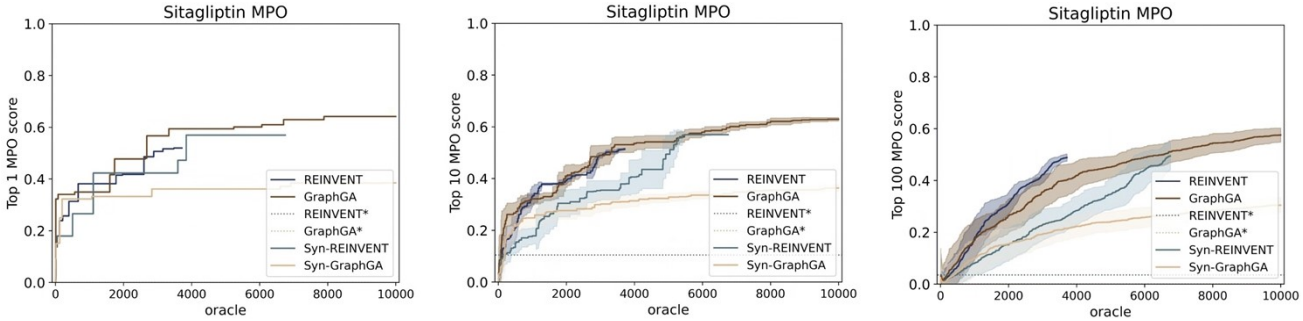


Fig. S24. The top-1 scores (right), top-10 scores (middle), and top-100 scores during the optimization process of REINVENT, GraphGA, Syn-REINVENT, Syn-GraphGA on Sitagliptin MPO.

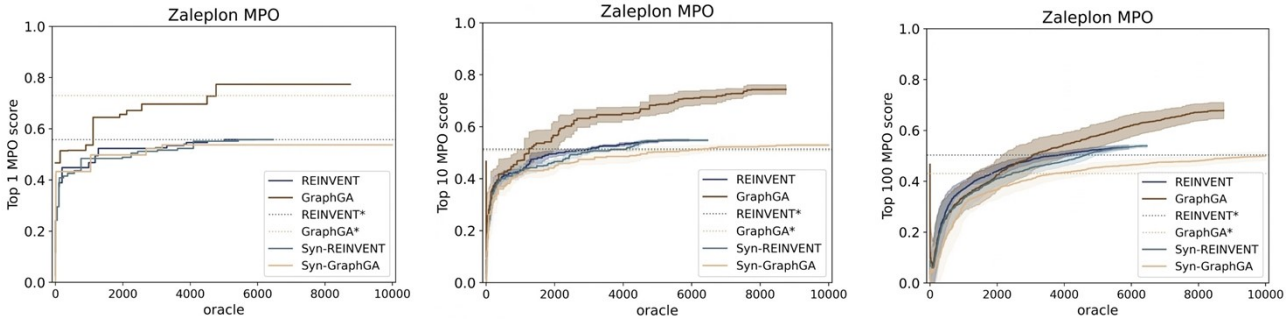


Fig. S25. The top-1 scores (right), top-10 scores (middle), and top-100 scores during the optimization process of REINVENT, GraphGA, Syn-REINVENT, Syn-GraphGA on Zaleplon MPO.

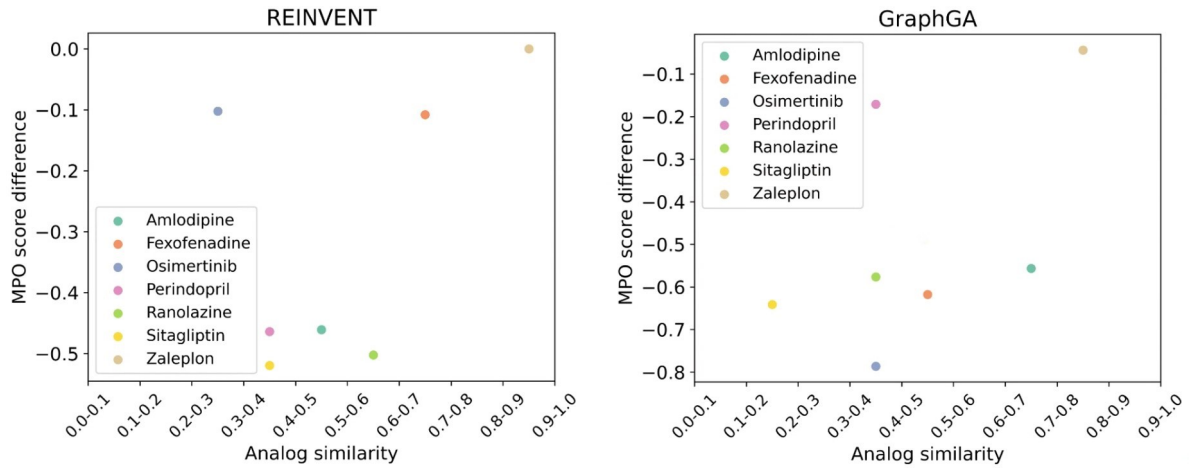


Fig. S26. The relationship between the analog similarity and score difference of the top-1 molecules generated by REINVENT and GraphGA.

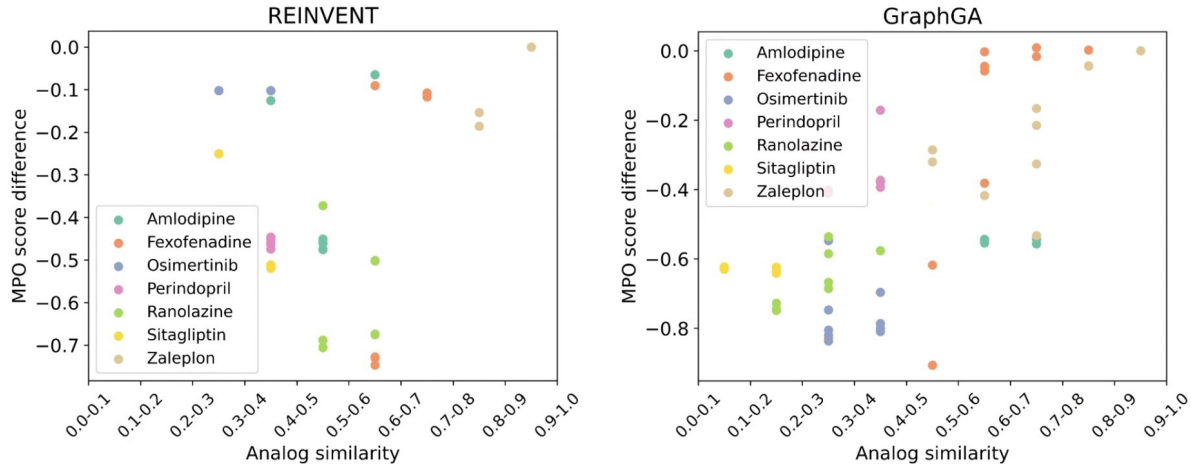


Fig. S27. The relationship between the analog similarity and score difference of the top-10 molecules generated by REINVENT and GraphGA.

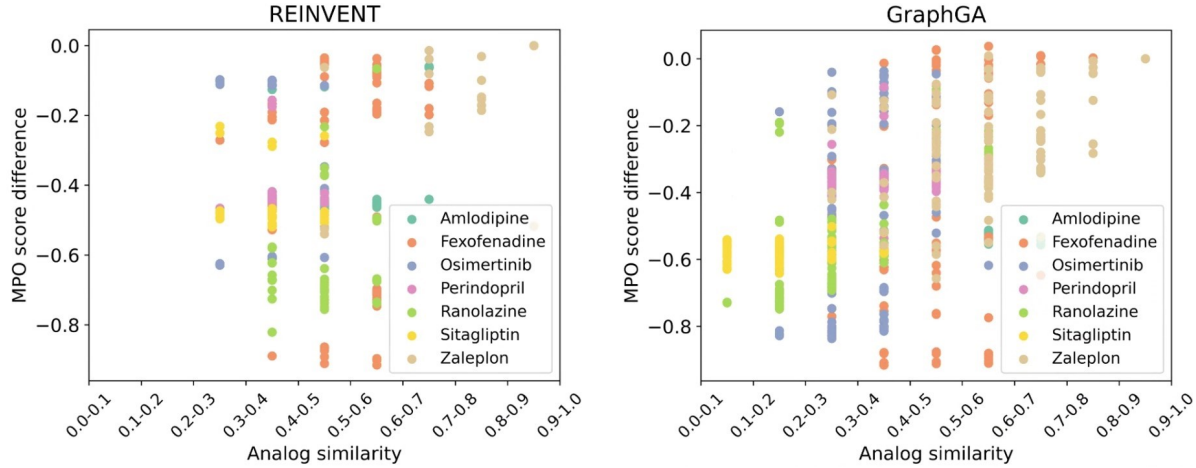


Fig. S28. The relationship between the analog similarity and score difference of the top-100 molecules generated by REINVENT and GraphGA.

S9. Inference time of SynTwins and baseline models

Table S3. The inference time and top-l average similarity of SynTwins, ChemProjector[10], and SynFormer[11].

	SynTwins	ChemProjector[10]	SynFormer[11]
Avg. speed @ CPU (s/mol)	66.1	73.2	83.4
Avg. Speed @ RTX3090 (s/mol)	-	12.6	10.2
Top-1 average similarity	0.8701	0.8018	0.6543
Top-3 average similarity	0.8209	0.7554	0.6247
Top-5 average similarity	0.7992	0.7268	0.6070

Reference

1 M. Hartenfeller, M. Eberle, P. Meier, C. Nieto-Oberhuber, K.-H. Altmann, G. Schneider, E. Jacoby and S. Renner, *J. Chem. Inf. Model.*, 2011, **51**, 3093–3098.
2 A. Button, D. Merk, J. A. Hiss and G. Schneider, *Nat Mach Intell*, 2019, **1**, 307–315.
3 C. W. Coley, W. H. Green and K. F. Jensen, *Journal of Chemical Information and Modeling*, 2019, **59**, 2529–2537.
4 The RDKit Documentation — The RDKit 2020.09.1 documentation, https://www.rdkit.org/docs/index.html.
5 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.
6 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
7 B. Chen, C. Li, H. Dai and L. Song, in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, 2020, pp. 1608–1616.
8 C. for D. E. and Research, Novel Drug Approvals at FDA, https://www.fda.gov/drugs/development-approval-process-drugs/novel-drug-approvals-fda, (accessed 5 March 2025).
9 W. Gao, R. Mercado and C. W. Coley, *arXiv*, 2022, preprint, arXiv:arXiv:2110.06389, DOI: 10.48550/arXiv.2110.06389.
10 S. Luo, W. Gao, Z. Wu, J. Peng, C. W. Coley and J. Ma, *arXiv*, 2024, preprint, arXiv:arXiv:2406.04628, DOI: 10.48550/arXiv.2406.04628.
11 W. Gao, S. Luo and C. W. Coley, *arXiv*, 2024, preprint, arXiv:arXiv:2410.03494, DOI: 10.48550/arXiv.2410.03494.
12 N. Brown, M. Fiscato, M. H. S. Segler and A. C. Vaucher, *J. Chem. Inf. Model.*, 2019, **59**, 1096–1108.
13 RDKit: Open-source cheminformatics, https://www.rdkit.org/, (accessed 15 February 2025).