Supporting Information
AI-Driven Synthetic Route Design Incorporated with Retrosynthesis Knowledge
Shoichi Ishidaa, Kei Terayamabc, Ryosuke Kojimad, Kiyosei Takasua, and Yasushi Okuno*cd

aa Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, 46-29 Yoshidashimo-Adachicho, Sakyoku, 606-8501 Kyoto, Japan
bb Graduate School of Medical Life Science, Yokohama City University, 1-7-29, Suehiro cho, Tsurumi ku, Yokohama 230 0045 Kanagawa, Japan
cc Graduate School of Medicine, Kyoto University, 53 Shogoin-Kawharacho, Sakyo-ku, 606-8507 Kyoto, Japan
dd HPC- and AI-driven Drug Development Platform Division, RIKEN Center for Computational Science, 7-1-26, Minatojima-minami-machi, Chuo-ku, Kobe 650-0047 Hyogo, Japan

Corresponding author
Please address the correspondence to Yasushi Okuno (okuno.yasushi.4c@kyoto-u.ac.jp)
Figure S1: Comparison of the numbers of solved molecules with ReTReK and ASCKOS. The gray bar with diagonal lines corresponds to ASCKOS with expansion time of 600 seconds. The gray, black, blue, orange, green, and red bars correspond to the no-knowledge, all-knowledge, STScore, CDSScore, ASScore, and RDScore patterns of ReTReK, respectively.
Figure S2: Comparison of the times necessary to solve compounds for different expansion sizes and retrosynthesis knowledge patterns. The gray, black, blue, orange, green, and red bars correspond to the no-knowledge, all-knowledge, STScore, CDSCore, ASScore, and RDScore patterns, respectively. The maximum reach of the whiskers in each boxplot is defined as 1.5IQR, where IQR represents the interquartile range. Outliers, defined as data points beyond the whiskers, are not shown in the boxplots.
Figure S3. All synthetic routes for three target compounds (a-b, ID: CHEMBL2068839, c-d, ID: CHEMBL4297623, and e, ID: CHEMBL3186534) found by ReTReK with retrosynthesis knowledge (a, c, and e) and a corresponding route found without retrosynthesis knowledge (b and d).
Figure S4. Solved synthetic routes for an EGFR kinase inhibitor\textsuperscript{1} using ReTReK with retrosynthesis knowledge. For the target compound considered here, synthetic routes were found using ReTReK with retrosynthesis knowledge, whereas no synthetic routes were found using the ReTReK without retrosynthesis knowledge.
Figure S5. Comparison of the synthetic route with each step's retrosynthesis knowledge score for three target compounds (a-b, hepatitis B virus capsid inhibitor, c-d, kwakhurin, and e-f, α7 nicotinic acetylcholine receptor silent agonist) found by ReTReK with retrosynthesis knowledge (a, c, and e) and the corresponding route found without retrosynthesis knowledge (b, d, and f). The first letter of each score (STScore, CDScore, ASScore, and RDScore) and the corresponding value are listed on each reaction arrow.
id: mol_15

id: mol_329
Figure S6 (pages ranging from S7 to S10). Solved synthetic routes for 10 molecules (mol_15, 329, 538, 591, 608, 616, 833, 845, 850, and 944) reported in Segler et al.\textsuperscript{2} using ReTReK with retrosynthesis knowledge. The expansion size is 500; the maximum number of iterations is set to 500; the weight parameters of the retrosynthesis knowledge scores, $w_1$, $w_2$, $w_3$, and $w_4$, are set to 5.0, 2.0, 0.5, and 2.0, respectively. In this condition, no synthetic route is found for the other nine molecules.
id: chem2018_1st_A_brд79_inhibitor

id: chem2018_1st_B_α-hydroxyetizolam

id: chem2018_1st_C_ATR_kinase_inhibitor
id: chem2018_1st_D_inhibitor_of_human_AML_cells

id: chem2018_2nd_A_S-4-hydroxyduloxetine

id: chem2018_2nd_B_5β-hydroxylurasidone
id: chem2018_2nd_C_dronedarone

id: chem2018_2nd_D_engelheptanoxide-C

id: chem2020_A_levomilnacipran
id: chem2020_D_imperanene

id: nature2020_tacamonidene
id: chem2018_1st_D_inhibitor_of_human_AML_cells

id: chem2018_2nd_A_S-4-hydroxyduloxetine
Figure S7 (pages ranging from S11 to S16). Solved synthetic routes for 11 molecules reported in 4 papers using Chematica/Synthia™ (BRD79 inhibitor, α-hydroxyetizolam, ATR kinase inhibitor, inhibitor of human AML cells, S-4-hydroxyduloxetine, 5β-hydroxylurasidone, dronedarone, engelheptanoxide-C, levomilnacipran, imperanene, and tacamonidine) using ReTReK. The expansion size is 500; the maximum number of iterations is set to 500; the weight parameters of the retrosynthesis knowledge scores, w1, w2, w3, and w4, are set to 5.0, 2.0, 0.5, and 2.0, respectively. In this condition, no synthetic route is found for the other 4 molecules (lycorane, aphanamal, dauricine, and lamellodysidine-A). As for the ASCKOS (pages ranging from S# to S#), Solved synthetic routes for 4 molecules (inhibitor of human AML cells, S-4-hydroxyduloxetine, dronedarone, and levomilnacipran) are shown. Two parameters of ASKCOS (the maximum number of templates and expansion time) were changed from the default setting. The maximum number of templates is equal to the expansion size of ReTReK and is set to 500. The expansion time is set to 600 seconds, which is the maximum time the application allows on their site (https://askcos.mit.edu/retro_interactive_mcts/).
Figure S8: Workflow of reaction template extraction. The reaction template extraction procedure consists of four steps. (1) Reaction records are standardized by removing explicit hydrogen, aromatizing, and keeping the largest fragments; (2) reaction records are narrowed down under the condition that a reaction is a single-step reaction that has a product and one to three reactants; (3) reaction templates are extracted from the filtered reaction records; (4) Sets of a product and the corresponding reaction template are filtered on condition that the reaction template can reversibly be applied to the product and derived reactants.
References

1. Su, Z.; Yang, T.; Wang, J.; Lai, M.; Tong, L.; Wumaier, G.; Chen, Z.; Li, S.; Li, H.; Xie, H.; Zhao, Z. Design, synthesis and biological evaluation of potent EGFR kinase inhibitors against 19D/T790M/C797S mutation. Bioorg. Med. Chem. Lett. 2020, 30, 127327. DOI: 10.1016/j.bmcl.2020.127327

2. Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018, 555, 604–610. DOI: 10.1038/nature25978