# Balancing the design of molecular structures with the design of their syntheses

Connor W. Coley

*Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States*

The description of the canonical DMTA cycle (design, make, test analyze) makes a strong distinction between the design of molecular structures and their synthesis. While moving from "design" to "make" does represent a transition from the virtual/computational world to the physical world, there is value in incorporating considerations of synthesizability into the molecular design process. Constraining oneself to in-stock or make-on-demand collections is one practical way to do this [1,2,3]. But if one wishes to leverage the 'creativity' of generative models, one runs into the issue that proposed molecular structures are often synthetically intractable [4]. Post hoc filtering with retrosynthetic planning programs (e.g., our own ASKCOS [5]) can be used to triage molecules, though this is rather inefficient. Fortunately, revisiting older ideas in reaction-based design originally applied to make-on-demand libraries allows one to devise deep learning approaches to synthesizability-constrained molecular design [6]. Such models can explore a superset of make-on-demand libraries when equipped with the same building blocks and transformation rules.

Beyond the generation of singleton structures, we and others have been exploring practical questions of *batched* molecular design. The ability to use parallel plate-based chemistry for library synthesis or common intermediates combined with diversification strategies saves synthetic cost on a per-candidate basis. Even relatively simple strategies grounded in cheminformatics like reaction pathway-constrained molecular generation can drive hit expansion efforts; hypothetical synthetic pathways can be scored in terms of their "diversifiability" and how fruitful a pathway-constrained enumeration might prove to be [7]. Finally, molecular designs come from a variety of sources in practice---from compound catalogs, to make-on-demand libraries, to generative models---and exhibit a wide variety of synthetic costs as a result. We have extended the framework of Bayesian optimization over molecules to account to quantitative balance the expected reward from a batch of molecules with the effort required to produce that batch [8]. We hope to continue down this line of research illustrating how quantitative, algorithmic decision-making can be used to drive design cycles in molecular discovery.

Bibliography :

[1] Graff, D. E., Shakhnovich, E. I., Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. Chem. Sci. 12, 7866-7881 (2021).
[2] Graff, D. E., Aldeghi, M., Marrone, J. A., Jordan, K. E., Pyzer-Knapp, E. O., Coley, C. W. "Self-focusing virtual screening with active design space pruning" arxiv: 01753 & J. Chem. Inf. Model. (2022).
[3] Fromer, J.C., Graff, D.E., Coley, C.W. "Pareto optimization to accelerate multi-objective virtual screening" arxiv:2310.10598 (2023).
[4] Gao, W., Coley, C. W. The synthesizability of molecules proposed by generative models. J. Chem. Inf. Model. 60(12) 5714–5723 (2020).
[5] https://askcos.mit.edu; https://gitlab.com/mlpds_mit/askcosv2/askcos-docs/-/wikis/home
[6] Gao, W., Mercado, R., Coley, C. W. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. ICLR [Spotlight] (2022).
[7] Levin, I., Fortunato, M.E., Tan, K.L., Coley, C.W. "Computer-aided evaluation and exploration of chemical spaces constrained by reaction pathways" AIChE J. DOI: 10.1002/aic.18234 (2023).
[8] Fromer, J., Coley, C.W. "An algorithmic framework for synthetic cost-aware decision making in molecular design" arxiv:2311.02187 (2023).