



Autonomous platforms for data-driven organic synthesis

Wenhao Gao¹, Priyanka Raghavan¹ & Connor W. Coley^{1,2}  

Achieving autonomous multi-step synthesis of novel molecular structures in chemical discovery processes is a goal shared by many researchers. In this Comment, we discuss key considerations of what an ideal platform may look like and the apparent state of the art. While most hardware challenges can be overcome with clever engineering, other challenges will require advances in both algorithms and data curation.

A framework for autonomous synthesis

In the iterative design, synthesis, and testing of new functional molecules, the rate at which candidate molecules can be physically realized often limits the rate at which computational designs can be validated. Platforms capable of performing chemical reactions in an automated or semi-automated manner, where the physical operations of a chemist are replaced by robotics and the planning by data-driven algorithms, can potentially mitigate this bottleneck.

The actualization of autonomous, data-driven organic synthesis will rely on advances in both *hardware* and *software* capabilities to overcome a combination of both *practical* and *scientific* challenges¹. In this Comment, we outline the major considerations that must be made to design autonomous platforms for target-oriented synthesis, progressing from the execution hardware, to synthesis planning, to adaptiveness and error handling, and finally to self-learning (Fig. 1).

Hardware requirements and desiderata

The basis of automated chemistry is the modularization of common physical operations to perform reactions: transferring a prescribed amount of starting materials to a reaction vessel, heating or cooling that vessel while mixing, purifying/isolating the desired product, analyzing the product, and using it in subsequent reaction steps. Fortunately, many requisite hardware units for such tasks have already been commercialized, such as liquid handling robots, robotic grippers for plate or vial transfer, computer-controlled heater/shaker blocks, and autosamplers for analytical instrumentation. Therefore, a straightforward (but not simple) paradigm for automated chemistry is to automate operations and sample transfer steps between existing lab hardware, exemplified by Burger et al.'s mobile robot chemist².

At the core of organic synthesis, automated reactions are run either in a flow or batch manner, with stirring, heating, and/or cooling capabilities. Key additional considerations when designing these automation components include minimizing evaporative losses, performing air-sensitive chemistries, and maintaining precise temperature control; all are addressable through engineering. An early platform, ChemKonzert, automated multi-step syntheses by replacing manual transfer operations with pumps but otherwise adhering to a typical batch process in round bottom flasks, separation flasks, and filters³—a paradigm since advanced and expanded upon by

¹Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ²Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ✉email: ccoley@mit.edu

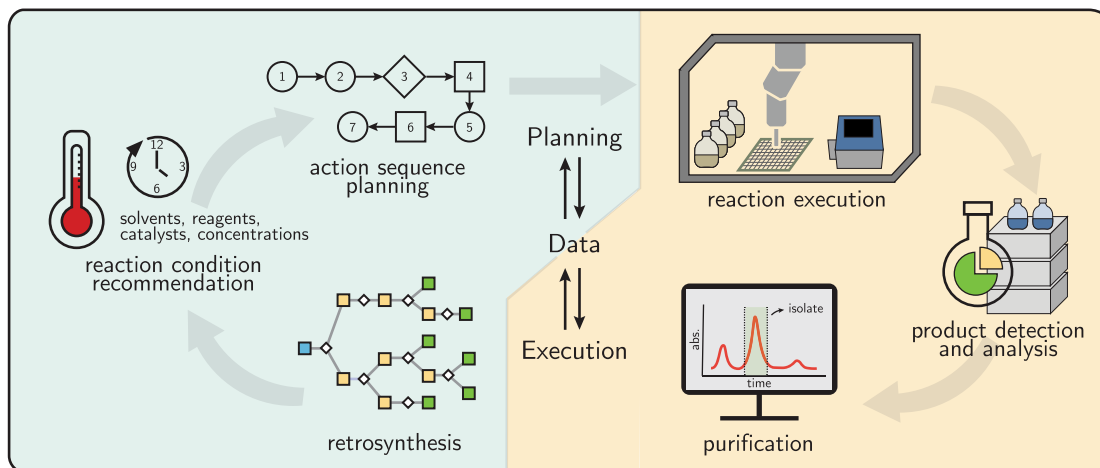


Fig. 1 A high-level workflow for autonomous data-driven organic synthesis. This process requires the close integration of synthesis planning and synthesis execution, linked by data that captures details of reaction and purification processes and their outcomes.

the Chemputer⁴. Additionally, from a practical standpoint, accessing a diverse chemical space requires equipping the platform with a suitably large chemical inventory of building blocks and reagents, otherwise these must be manually prepared prior to any synthesis. In medicinal chemistry applications, Eli Lilly has been a leader in automated multi-step synthesis by designing a platform around microwave vials as reaction vessels⁵, with significant ongoing investments in automation more broadly including a chemical inventory able to store five million compounds⁶. Flow platforms too can be automated for target-oriented synthesis using computer-controlled pumps and flow-path reconfiguration^{7,8}, though they require additional planning considerations (e.g., solubility).

Following the reaction, liquid chromatography–mass spectrometry (LC/MS) is most commonly used for analysis or quantitation. Multi-step reactions add a layer of complexity, as crude products must be isolated and resuspended in solvent between reactions. This invites new challenges with regards to the automation of solution transfer between the reaction area and the purification and analysis unit. Constraining the reaction space to a specific subset can mitigate the burden of purification as exemplified by Burke’s iterative MIDA-boronate coupling platform that uses a catch and release method applicable to a specific reaction;⁹ however, a universally applicable purification strategy does not yet exist.

Synthetic planning including and beyond retrosynthesis

At a high-level, an autonomous chemistry platform must decide what it *intends* to do (as a chemist might describe it) and then translate that into what it should *actually* do (in terms of physical operations). The latter step is dependent on available hardware operations, but can leverage protocols intended to be hardware-agnostic, such as the chemical description language (XDL)¹⁰. When targeting structures without known experimental procedures, the former step requires developing software tools for computer-aided synthesis planning, including and beyond retrosynthesis.

Computer-aided retrosynthesis has arguably existed almost since the concept of retrosynthesis itself, but has failed to gain traction due to a perception that proposed routes are of low quality. As a turning point, Segler et al.¹¹ pioneered a data-driven approach using a Monte Carlo tree search that passed a “chemical Turing test”, wherein graduate-level organic chemists expressed

no statistically significant preference between literature-reported routes and the program’s. Mikulak-Klucznik et al.¹² further demonstrated that this approach is viable for complex natural products with their expert (not data-driven) program, Synthia. Their successes have catalyzed an interest in developing neural models that learn allowable chemical transformations from reaction databases. These models are commonly divided into template-based and template-free approaches depending on whether the model makes use of symbolic pattern-matching rules. Both types have been incorporated with semi-automated synthesis platforms to streamline target-oriented organic synthesis, for example, with Coley et al.’s ASKCOS⁷ and IBM’s integration with commercial hardware¹³. However, successful applications of data-driven retrosynthesis with automation have been relatively simple molecules, where few (1–5) steps are required and where stereocenters are typically sourced from building blocks rather than installed.

It is essential to recognize that retrosynthesis is merely the first step of autonomous organic synthesis, as it does not address numerous practical considerations. Experimental execution of a synthetic route requires specification of quantitative reaction conditions—amounts of each reactant, solvent(s), temperature, time, etc.—and how this translates into a detailed action sequence for the hardware to follow, at the very least specifying order of addition. Subtle changes in procedure can significantly affect the reaction outcomes, but these subtleties are missing from current databases, and therefore are also missing from current data-driven tools¹⁴. Proposed synthetic routes also require further scoring and ranking in terms of their automation compatibility, perhaps tailored to a specific hardware platform, not just chemical feasibility. These steps remain largely unaddressed.

Error handling and robustness to mispredictions

The need for precise planning can be partially mitigated with platforms that can cope with mispredictions and are able to adaptively determine a suitable action sequence through trial and error. The optimization of reaction yields or selectivities by modulating reaction conditions is a decades-old task, with recent demonstrations including applications of statistical optimization to multi-step flow chemistry¹⁵, as well as applications of Bayesian optimization¹⁶. Predicted conditions may be suitable as an initial guess to be further improved through empirical optimization. This workflow of incorporating reaction screening and

optimization between synthesis planning and multi-step synthesis is exemplified by SRI's SynFini⁸.

However, even this basic level of adaptivity relies on a number of capabilities that are not trivial to automate, such as confirming product identities and quantifying their yields without relying on a user-provided product standard or calibration curve. Most platforms are currently equipped with only LC/MS, while structural elucidation or quantitation may require instruments such as nuclear magnetic resonance (NMR) or corona aerosol detection (CAD), the latter of which promises to enable universal calibration curves. While the hardware challenges can be overcome through engineering efforts, the computer-assisted structural elucidation and prediction of analytical response factors would benefit from renewed attention.

"Failures" in autonomous target-oriented synthesis will often be more dire than a subpar yield. Chemistry is sufficiently complex that predictive models might never be perfectly predictive of physical reality; moreover, we may wish to explore *new* reactivity, which by some definitions is inherently less predictable. A key reaction step might not produce any desired product, warranting a complete revision of that route to circumvent that false-positive prediction. Flow chemistry platforms may be prone to clogging and therefore necessitate a means to detect and recover from such events. Plate-based or vial-based platforms are in principle more robust, provided that the reaction vessel is disposable and can simply be discarded if the procedure fails.

Self-learning and improvement

Beyond handling errors during synthesis, an ideal autonomous platform would learn and improve over time just as a chemist accrues knowledge and experience throughout their career. Arguably, these attributes of continual learning and the ability to respond to unforeseen outcomes are what would make a platform autonomous, rather than merely automated.

There are at least two factors that complicate this goal of life-long learning for data-driven platforms. First, the volume of data generated by a single platform will be overshadowed by the historical reactions tabulated in reaction databases; for new data to influence predictions, it may need to be treated separately by the algorithms (e.g., as a fine-tuning set) rather than integrated with a broader knowledge base. Second, the type of generated data will be qualitatively different from existing databases: it has the potential to be far richer in terms of procedural details and analytical chemistry, but will likely be unable to match the substrate diversity of published reactions given a practically sized chemical inventory. How to leverage this multi-modality is a new challenge in algorithm design.

Outlook

Many elements of an autonomous platform for data-driven organic synthesis exist, yet we will continue to be stuck in the proof-of-concept phase unless several shortcomings are resolved. The level of precision required to execute synthetic pathways is not matched by the planning algorithms, and key challenges such as purification design remain almost entirely unaddressed. Data availability is a particular impediment, although nascent efforts like the Open Reaction Database¹⁷ may address this in time. Transitioning from "automation" to "autonomy" implies a certain degree of adaptiveness that is difficult to achieve with the limited analytical capabilities of many platforms. While already useful in isolation, these platforms will be particularly enabling when integrated with molecular design algorithms for *function-oriented*

synthesis. In this setting, one can reconsider the role these platforms are meant to play: the ability to achieve any target molecular function may be more important than the ability to achieve any target molecular structure, which could make certain platform limitations (e.g., in terms of scope of reaction types) perfectly acceptable.

Received: 14 December 2021; Accepted: 8 February 2022;
Published online: 28 February 2022

References

1. Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous discovery in the chemical sciences part II: outlook. *Angew. Chem. Int. Ed.* **59**, 23414–23436 (2020).
2. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
3. Doi, T. et al. A formal total synthesis of taxol aided by an automated synthesizer. *Chem. Asian J.* **1**, 370–383 (2006).
4. Steiner, S. et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, eaav2211 (2019).
5. Godfrey, A. G., Masquelin, T. & Hemmerle, H. A remote-controlled adaptive medchem lab: an innovative approach to enable drug discovery in the 21st century. *Drug Discov. Today* **18**, 795–802 (2013).
6. Eli Lilly and Company in Collaboration with Strateos, Inc. Launch Remote-Controlled Robotic Cloud Lab | Eli Lilly and Company. <https://investor.lilly.com/news-releases/news-release-details/eli-lilly-and-company-collaboration-strateos-inc-launch-remote> (accessed 15 October 2021).
7. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
8. Collins, N. et al. Fully automated chemical synthesis: toward the universal synthesizer. *Org. Process Res. Dev.* **24**, 2064–2077 (2020).
9. Li, J. et al. Synthesis of many different types of organic small molecules using one automated process. *Science* **347**, 1221–1226 (2015).
10. Angelone, D. et al. Convergence of multiple synthetic paradigms in a universally programmable chemical synthesis machine. *Nat. Chem.* **13**, 63–69 (2021).
11. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
12. Mikulak-Klucznik, B. et al. Computational planning of the synthesis of complex natural products. *Nature* **2020**, 1–11. <https://doi.org/10.1038/s41586-020-2855-y> (2020).
13. IBM RXN for Chemistry. <https://rxn.res.ibm.com/rxn/robo-rxn/welcome> (accessed 15 October 2021).
14. Gao, H. et al. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).
15. Bédard, A.-C. et al. Reconfigurable system for automated optimization of diverse chemical reactions. *Science* **361**, 1220–1225 (2018).
16. Shields, B. J. et al. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).
17. Kearnes, S. M. et al. The open reaction database. *J. Am. Chem. Soc.* <https://doi.org/10.1021/jacs.1c09820> (2021).

Acknowledgements

The authors thank the Machine Learning for Pharmaceutical Discovery Synthesis Consortium, the MIT Research Support Committee, and the Office of Naval Research under grant number N00014-21-1-2195 for financial support.

Author contributions

W.G., P.R., and C.W.C. all contributed to the writing of the article; C.W.C. prepared the figure.

Competing interests

C.W.C. is a scientific advisor to and shareholder of companies including Entos, Inc. and Kebotix. W.G. and P.R. declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Connor W. Coley.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022