https://doi.org/10.1038/s44160-023-00335-1

Automation, analytics and artificial intelligence for chemical synthesis

Junliang Liu & Jason E. Hein

Check for updates

Automation and real-time reaction monitoring have enabled data-rich experimentation, which is critically important in navigating the complexities of chemical synthesis. Linking real-time analysis with machine learning and artificial intelligence tools provides the opportunity to accelerate the identification of optimal reaction conditions and facilitate error-free autonomous synthesis. This Comment provides a viewpoint underscoring the growing significance of data-rich experiments and interdisciplinary approaches in driving future progress in synthetic chemistry.

The appearance of SARS-CoV-2 in 2019 set many ground-breaking changes in motion. The urgent and immediate threat to human health presented the global scientific community with one of the most significant challenges of our generation. Even in the earliest days of the virus' spread, synthetic chemistry was poised to play an enormous role in mitigating the global pandemic[1]. One of the most striking advances that has emerged from these circumstances was the development of nirmatrelvir; an orally bioavailable protease inhibitor developed by Pfizer. This drug development campaign set the record for the fastest commercial development of a novel pharmaceutical, moving from small-scale discovery, though the gauntlet of preclinical toxicology, phase I–III trials and gaining emergency use authorization in only 17 months[2]. Such fast development at a large scale required the concerted efforts of a multitude of scientific disciplines, with synthetic chemistry featuring prominently. The synthetic teams had to identify optimal feedstock materials, solvents, reagents, and catalysts to ensure consistent access to material with exceptionally limited time. The nirmatrelvir project highlights the critical role reaction process data holds in rapid and successful decision making.

The role of automation when navigating the synthesis maze

Synthesizing most molecules requires navigating a multi-step transformation, balancing input materials (solvents, reagents, catalysts), reaction parameters (temperatures, order of additions, time) as well as workup and purification strategies. Traversing this multifactorial challenge is analogous to searching through a maze with limited resources. Historically, chemists had to draw on prior experiences, create careful strategies and make decisions with limited data. Enabling technology, such as laboratory automation, radically changed the landscape, enhancing both the quantity and accuracy of analytical reaction data, allowing better decisions in less time. Techniques, such as high-throughput experimentation (HTE) can be deployed to rapidly survey possible reaction conditions[3], but these techniques usually only provide an analytical percent yield at a fixed reaction time, forfeiting critical details pertaining to the reaction mechanism or dynamics (Fig. 1).

The corollary strategy, termed data-rich experimentation (DRE), focuses on extracting real-time reaction progress data, quantifying all measurable species or parameters, and providing a comprehensive play-by-play for a single reaction. Route scouting and optimizing using real-time monitoring provides a detailed picture of the reaction kinetics, revealing critical information such as reaction intermediates, rate constants, and by-product reaction pathways. Automation is the key enabling tool to make DRE approaches a manageable and productive endeavour. First, hardware and instrument automation are needed to accurately orchestrate the capture and analysis of reaction aliquots repeatedly over the entire reaction progression. This applies to different degrees depending on the analytical technology of choice (high-performance liquid chromatography–mass spectrometry, nuclear magnetic resonance spectroscopy, high-resolution mass spectrometry) but in all cases the frequency, precision and extended duration demanded of the reaction progress measurement disincentivizes manual operation. Second, the sheer volume and complexity of reaction analytical data requires software automation techniques to help annotate, process, and convert the raw data into trends representing concentration versus time arrays for each reaction component. Finally, complex reaction manipulations can be executed with automation, such as precise variations in temperature or catalyst dosing, allowing data to be extracted from a single reaction, which would typically require multiple experiments.

Replacing clockwork executing with 'rules' and 'goals'

The current paradigm of data-driven reaction investigation focuses almost exclusively on using human-in-the-loop steps for converting data to information. This means the analytical tools create real-time reaction trends that are then interpreted by an operator to plan or guide the experimental campaign. Fixed multivariant statistical tools, such as design of experiments or optimization strategies such as batch-Bayesian optimization leverage automation to acquire large data sets, but the final interpretation and scripting is manual.

An emergent opportunity now exists, where telemetry from real-time monitoring can be used to dramatically accelerate process optimization and reaction discovery. Real-time data can be leveraged, enabling automated systems to receive critical feedback on the process. This both ensures accurate execution of the intended experiment and enhances the transferability and reproducibility of the automated synthetic protocol. The same data set can be used to allow the automated reaction hardware to adapt to variable circumstances.

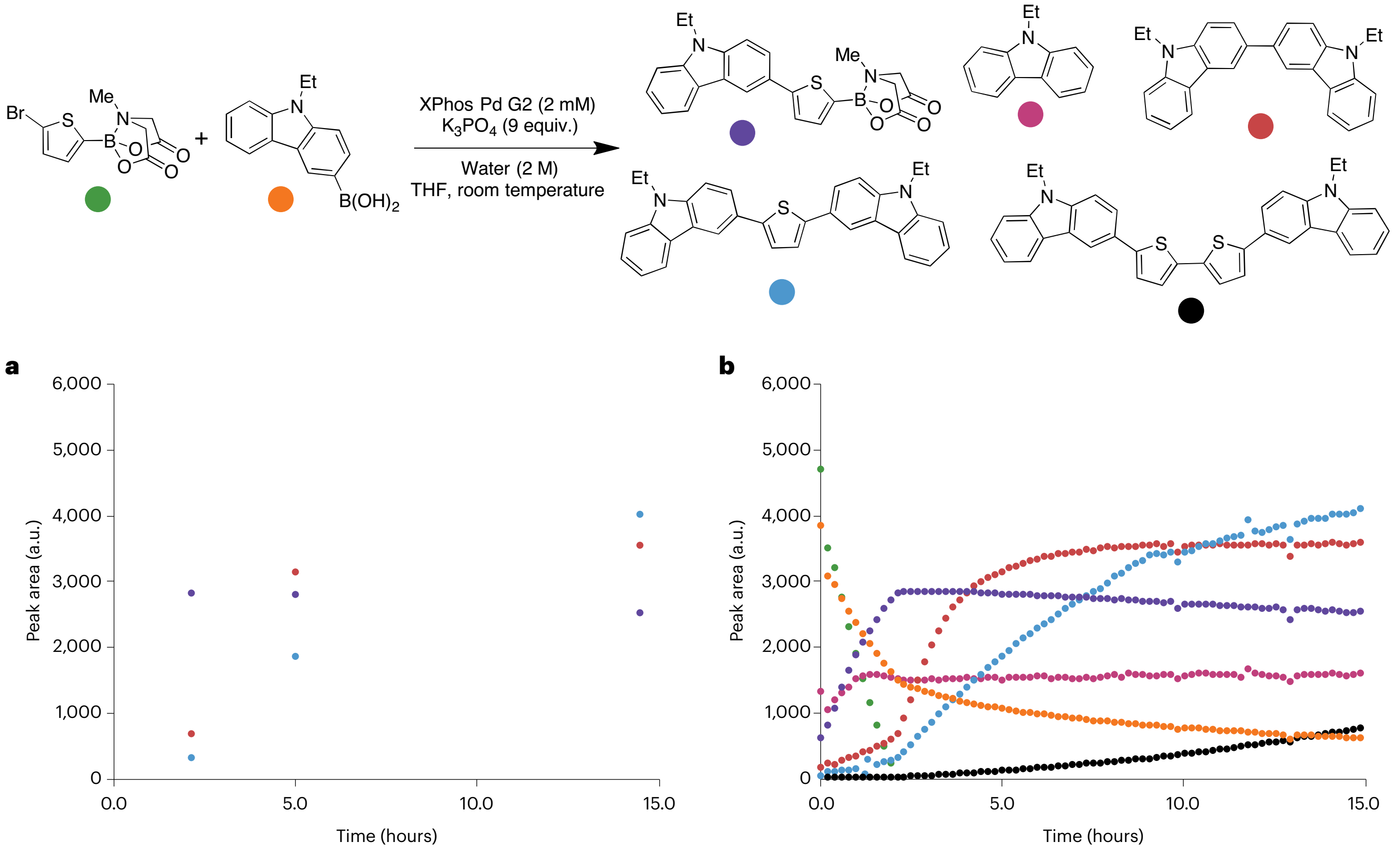

Fig. 1 | Suzuki–Miyaura cross-coupling analysed by ultra-high-performance liquid chromatography showing the different peak areas versus time for the starting materials as well as products and common by-products. a, Limited understanding is achieved when few time-points are captured. **b**, The identical transformation visualized as the full reaction profile immediately provides a comprehensive view for the reaction. a.u., arbitrary units; XPhos Pd G2, chloro(2-dicyclohexylphosphino-2′,4′,6′-triisopropyl-1,1′-biphenyl)[2-(2′-amino-1,1′-biphenyl)]palladium(II); THF, tetrahydrofuran.

This is incredibly impactful when trying to execute multi-step transformations, where a precursor must be formed prior to the synthesis proceeding. In place of a hardcoded script that would dose a fixed quantity of material at a set time, the reactor can be trained to add enough reagent when the first reaction is finished[4]. These conditional arguments allow on-the-fly corrections more typical of traditional research and development workflows and open the door for error-free autonomous synthesis by providing a synthetic 'goal' to be achieved following experimental 'rules'.

The potential of AI- and ML-enhanced reaction design

Machine learning (ML) and artificial intelligence (AI) tools are powerful additions to experimental data-driven workflows for accelerating the identification of reaction conditions[5]. Predictive models have been built from experimental data obtained from HTE or literature sources, which can suggest reaction conditions to execute an unknown transformation[6]. In addition, autonomous optimization platforms have been created by fusing robotic reaction execution, end-point sampling and data extraction with ML optimization algorithms[7]. Using these approaches, it is possible to reduce the number of experiments required to identify the ideal conditions, however, both examples reduce the experimental outcome to a single score of quantity, such as percentage yield or percentage stereoselectivity[8]. These strategies have merit, but reduction to a single measurement at a fixed time belies the complexity inherent to chemical reactions.

Many studies have demonstrated that drawing reaction performance data (yield) from existing literature leads to mixed results. Data are biased towards most frequently published conditions, leading to extraction of popular reaction parameters rather than optimal conditions[9]. Worse yet, the heterogeneity in both the quantitative measurements, and conditions or techniques applied make it impossible to distinguish if a reported yield is the result of an experimental failure or difficulty in isolation. Attempts to homogenize and systematize reported synthetic data are emerging, however, they are yet in their infancy.

Data sets generated from HTE automation systems are more consistent, however may still provide systemic bias, limiting their broad applicability. In particular, the time point chosen to assay for the chosen analytical metric may deliver false-positive, or -negative data. For example, a low recorded product yield could be due to a reaction combination that had a delayed initiation, or if the desired product was unstable under the reaction conditions. Thus, choosing the wrong time window to query the reaction can lead to dramatic oversimplification or misinterpretation of the system under interrogation. While sparse data from HTE can act as a guidepost, many truly interesting and unexpected breakthroughs are missed.

Real-time reaction monitoring presents a critical advantage, whereby predictive models could be trained using the full kinetic data. These comprehensive data address all issues relating to data integrity, bias and oversimplification. First, by recording the entire reaction profile, variations in reaction performance by different researchers could be captured and explained. Mismatches would serve to focus efforts to rectify issues of failed transfer of a protocol. Second, the full evolution of reaction species would be captured, allowing the evolution of the target material to be delineated, as well as by-products and intermediates. These trends would serve as useful metadata for future reaction discovery as they capture transformations that are possible even if they are not the focus of the study. Finally, very few reaction trends may be required to unambiguously classify the underlying mechanism using an appropriately trained neural network[10]. In general, the pattern recognition ability of ML-methods is well suited to train on the complex pattern from the entire reaction.

Overall, the data-science revolution in synthetic chemistry is accelerating, enhancing the need for robust, data-rich experiments. Real-time reaction analytics have already been leveraged to dramatically reduce the time needed to reach a molecular target. By further linking these automated data-gathering methods with new ML and AI tools, our ability to predict optimal conditions and discover new synthetic routes will grow exponentially. It is through these new interdisciplinary approaches that the record-breaking pace for the commercialization of nirmatrelvir will become business as usual.

Junliang Liu[1] & Jason E. Hein[1,2,3]
[1]Department of Chemistry, University of British Columbia, Vancouver, British Columbia, Canada. [2]Acceleration Consortium, University of Toronto, Toronto, Ontario, Canada. [3]Department of Chemistry, University of Bergen, Bergen, Norway.
e-mail: jhein@chem.ubc.ca

Published online: 30 May 2023

References

1. Hardy, M. A. et al. *ACS Cent. Sci.* **6**, 1017–1030 (2020).
2. Allais, C. et al. *ACS Cent. Sci.* https://doi.org/10.1021/acscentsci.3c00145 (2023).
3. Stevens, J. M. et al. *Org. Process Res. Dev.* **26**, 1174–1183 (2022).
4. Liu, J., Sato, Y., Yang, F., Kukor, A. J. & Hein, J. E. *Chem. Methods* **2**, e202200009 (2022).
5. Shi, Y., Prieto, P. L., Zepel, T., Grunert, S. & Hein, J. E. *Acc. Chem. Res.* **54**, 546–555 (2021).
6. Shields, B. J. et al. *Nature* **590**, 89–96 (2021).
7. Christensen, M. et al. *Commun. Chem.* **4**, 112 (2021).
8. Zahrt, A. F. et al. *Science* **363**, eaau5631 (2019).
9. Beker, W. et al. *J. Am. Chem. Soc.* **144**, 4819–4827 (2022).
10. Burés, J. & Larrosa, I. *Nature* **613**, 689–695 (2023).

Acknowledgements

The authors' research team is supported by the Canada Foundation for Innovation (CFI-35883) and the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2021-03168). This work is dedicated to Prof. Philip G. Hultin (1961–2018).

Competing interests

The authors declare no competing interests.