AI, AUTOMATION, AND ROBOTICS TRANSFORM CHEMICAL DISCOVERY AND DEVELOPMENT

KLAVS F. JENSEN*

*Chemical Engineering and Materials Science & Engineering, MIT, Cambridge, MA 02139, U.S.A.

Introduction

Artificial Intelligence (AI), specifically machine learning (ML) tools, is becoming increasingly adept at generating new candidate molecules, predicting their properties, proposing reaction pathways using computer-aided synthesis planning (CASP), and aiding in the analysis of analytical data [1]. Automation and robotic technologies have also become easier to use and more affordable to integrate, enabling automated chemical synthesis and characterization with little or no human intervention once the system is set up [2]. These self-driving laboratory systems typically involve multiple stationary robots that prepare and transfer samples through the various stages of synthesis and characterization. The robots can also roam through the lab as they transfer samples between process units, analogous to a human operator [3]. The ability to purify and measure reaction outcomes is essential to automated systems to learn and optimize performance. Feedback based on reaction conditions and measured outcomes enables the optimization of selected performance metrics across continuous and categorical variables by using a range of optimization techniques, now predominantly Bayesian methods [2].

Integrating these automated synthesis systems with ML algorithms for molecular generation [4], property prediction [5], computer-aided synthesis (CASP) [6, 7], and ML-based chemical analysis [8] into the traditional design-make-test-analyze (DMTA) workflow has advanced the development of autonomous chemical discovery platforms capable of functioning across diverse chemical spaces with minimal manual intervention [9]. However, human operators will continue to play a crucial role in operating these platforms by defining goals, initializing procedures, monitoring experiments, assisting with error recovery, and managing resources.

Property–focused discovery platforms can propose and synthesize molecules to expand the training datasets for ML generative and property prediction models, helping to map the chemical space and ultimately identify top-performing molecules. Once established, a platform can be adapted for other applications by modifying the underlying ML models. For instance, a system initially designed to discover new organic dye molecules [9] was

repurposed for drug discovery by incorporating additional assays, ML property models, different ML molecular generative models, and Bayesian optimization to balance experimental costs and the number of molecules screened during sequential rounds of virtual, coarse, and refined experimentation [10].

Large Language Models (LLMs) are likely to accelerate further the integration of AI tools into automated chemical experimentation and analysis by helping researchers streamline and enhance their workflows through intuitive natural language prompts [11-14]. In a sense, LLMs act as helper tools that enable access to advanced ML computational resources, making data-driven methodologies more accessible to a wider community of chemists and facilitating straightforward interfaces with synthesis, purification, and analysis equipment through a unified interface.

Our recent research contributions

Our research has focused on advances in chemistry, engineering, and machine learning that are necessary for automated and accelerated chemical discovery and development. For this session, two recent studies [9, 10] towards autonomous platforms for property-driven molecular discovery serve to exemplify challenges and opportunities in integrating automation and ML techniques in chemical synthesis of organic molecules. The first case study uses the discovery of new organic dyes as a test case, as their fundamental properties — absorption maximum, water-octanol partition coefficient, and photo-oxidative stability — are readily measurable, and their realization involves a multi-step synthesis through a rich variety of chemical transformations. The platform's master controller orchestrates automation and ML prediction tools to iteratively propose, realize, and characterize dye molecules within the DMTA cycle.

Initially, a generative ML model created candidate molecules. We selected a graph completion approach that decorates molecular scaffolds to reduce the risk of generating unstable or non-synthesizable structures, a common issue for generative models [15]. For each generated candidate, multiple synthesis pathways were automatically planned and proposed using open-source retrosynthesis ML tools, ASKCOS [6]. Approximately 10-20% of these molecules had retrosynthesis routes that ended in purchasable starting materials. Most reaction pathways for the generated molecules required several steps, giving access to a broader range of the multidimensional property space.

To enable the system to execute the identified reaction routes, predicting reaction conditions (e.g., reagents, catalysts, solvents, equivalence ratios, and concentrations) with an ML component of ASKCOS was essential. However, predicting these conditions remains a major challenge due to limited accurate data, requiring the use of checks and heuristic rules to supplement ML predictions. Only recently have models shown improved performance over chemical intuition and nearest neighbor approaches [16].

Ideally, it would also be useful to predict expected yields for the reactions; however, yield prediction is a notoriously challenging problem, with only limited success for large Suzuki [17] and Buchwald-Hartwig coupling reaction datasets [18].

The properties of the proposed molecules were evaluated using ML molecular prediction models [5], and the feasible candidate pathways were scored based on their molecular value and platform feasibility. The platform-selected reaction pathways were automatically translated into synthesis and characterization workflows to be executed in 96-well plates. The master controller orchestrated four independent systems with different capabilities to work simultaneously, executing reactions, preparing reaction solutions, analyzing reaction outcomes, isolating target products, and characterizing the isolated molecules. The developed platform was capable of performing multiple unrelated tasks in parallel and only required human intervention for non-automated error recovery and restocking. We developed our own software after considering alternative options, such as Chemputer [19] and ChemOS [20], since none of these options supported parallel operation in well plates, work-up and isolation, characterization, and on-the-fly modification of workflows. The development of open-source, flexible, user-friendly orchestration and operation software remains a challenge.

After completing reaction tasks, a series of automatically selected and executed work-up steps processes the crude products to prepare them for subsequent reactions and HPLC analysis. We employed multivariate curve resolution and a photodiode array (PDA) detector to deconvolute peaks, in conjunction with a mass spectrometer (MS) to identify them, and an ML model of molar extinction coefficient to determine analyte concentration without calibration [8]. Purification often presents a challenge for automated systems, as reactions typically require cleanup through extraction, evaporation, precipitation, and filtration. Additionally, chromatography demands method development to handle diverse chemical compositions. Vendor-proprietary software protocols also create a significant bottleneck for achieving fully autonomous systems. High-resolution NMR is typically required for structural characterization; however, for most laboratories, it remains too costly to incorporate into an automated reaction platform. Mobile robots are one option that allows for automatic transfer to centralized NMR facilities [3] - we hand-carried our samples.

A plate reader measured absorption spectra, calibrated HPLC retention times provided water/octanol partition coefficients, and a simulated solar light source combined with the plate reader quantified photo-oxidative degradation. The measured molecular properties were automatically fed back to retrain the property prediction models, completing one step of the automated DMTA cycle. Three iterations were sufficient for the ML model deviations to approach the experimental uncertainty, allowing for exploitation. Human involvement was limited to setting and adjusting objectives, providing requested

materials, and occasionally fixing unrecoverable errors, such as clogging of the HPLC unit. Overall, the platform attempted over 3,000 reactions, with more than 1,000 yielding the predicted reaction product, thereby completing multi-step reaction pathways for 318 previously unreported molecules, demonstrating its ability to explore unknown structure-property spaces by searching for structures with desired properties (hits) and to exploit characterized structure-property spaces by optimizing promising candidates (leads).

Since our system was constructed with standard equipment modules, it could be easily adapted to an autonomous chemical synthesis and testing platform for automatically searching for new histone deacetylase inhibitors by incorporating new assays and ML tools [10]. The new platform combined a genetic algorithm to generate diverse candidate molecules, a multi-fidelity Bayesian optimization iterative discovery algorithm for molecular property optimization to select candidates and determine the optimal level of fidelity at which to evaluate them, and ASKCOS computer-aided synthesis planning tools to plan synthesis execution.

Challenges and Outlook

Self-driving laboratory systems for automated discovery and development with varying degrees of autonomy will become as ubiquitous in chemical laboratories as today's HPLCs. They will accelerate discovery, expand chemical data, and drive innovation. However, many challenges remain to improve the accuracy of ML models and ease the integration of equipment. The cost and complexity of such systems are barriers that can be mitigated by utilizing standardized and scalable equipment modules, as well as open-source software. Their effective operation requires the collaboration of chemists, computer scientists, and engineers who have some understanding of each other's disciplines. Advanced models, such as large language models (LLMs), are poised to play an increasingly important role in experimental workflows due to their ability to predict properties, synthesize new molecules, and orchestrate existing computational and experimental tools through a unified interface. However, challenges remain, including minimizing hallucinations, improving data efficiency in training, and increasing integration with automation and robotics.

FAIR (Findable, Accessible, Interoperable, Reusable) data practices will be essential for improving the ML predictions of retrosynthesis tools, particularly in predicting reaction conditions and reactivity. The organic chemistry database ORD is an example of a community-driven effort to create access to reliable, high-quality reaction data [21]. The autonomous systems described in the main text automatically saved all reaction information, including conditions and outcomes for both failed and successful reactions, for all attempted reactions, illustrating that current and future self-driving systems offer the opportunity to create large, community-accessible databases for training improved ML models.

Acknowledgments

The author thanks his MIT colleagues and lab members for the collaborations underlying the discussed research, in particular Drs. Canty, Koscher, and McDonald.

References

1. A.M. Mroz, A.R. Basford, F. Hastedt, I.S. Jayasekera, I. Mosquera-Lois*, et al.*, *Chemical Society Reviews*, **54**, 5433 (2025).
2. G. Tom, S.P. Schmid, S.G. Baird, Y. Cao, K. Darvish*, et al.*, *Chemical Reviews*, **124**, 9633 (2024).
3. T. Dai, S. Vijayakrishnan, F.T. Szczypiński, J.-F. Ayme, E. Simaei*, et al.*, *Nature*, **635**, 890 (2024).
4. C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, and K.F. Jensen, *WIREs Computational Molecular Science*, **12**, e1608 (2022).
5. E. Heid, K.P. Greenman, Y. Chung, S.-C. Li, D.E. Graff*, et al.*, *Journal of Chemical Information and Modeling*, **64**, 9 (2024).
6. Z. Tu, S.J. Choure, M.H. Fong, J. Roh, I. Levin*, et al.*, *Accounts of Chemical Research*, **58**, 1764 (2025).
7. B. Mikulak-Klucznik, P. Gołębiowska, A.A. Bayly, O. Popik, T. Klucznik*, et al.*, *Nature*, **588**, 83 (2020).
8. M.A. McDonald, B.A. Koscher, R.B. Canty, and K.F. Jensen, *Chemical Science*, **15**, 10092 (2024).
9. B.A. Koscher, R.B. Canty, M.A. McDonald, K.P. Greenman, C.J. McGill*, et al.*, *Science*, **382**, eadi1407 (2023).
10. M.A. McDonald, B.A. Koscher, R.B. Canty, J. Zhang, A. Ning*, et al.*, *ACS Central Science*, **11**, 346 (2025).
11. A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A.D. White*, et al.*, *Nature Machine Intelligence*, **6**, 525 (2024).
12. D.A. Boiko, R. Macknight, B. Kline, and G. Gomes, *Nature*, **624**, 570 (2023).
13. Z. Zheng, O. Zhang, H.L. Nguyen, N. Rampal, A.H. Alawadhi*, et al.*, *ACS Central Science*, **9**, 2161 (2023).
14. M.C. Ramos, C.J. Collison, and A.D. White, *Chemical Science*, **16**, 2514 (2025).
15. W. Gao and C.W. Coley, *Journal of Chemical Information and Modeling*, **60**, 5714 (2020).
16. X. Sun, J. Liu, B. Mahjour, K.F. Jensen, and C.W. Coley, *Chemical Science*, (2025).
17. P. Raghavan, A.J. Rago, P. Verma, M.M. Hassan, G.M. Goshu*, et al.*, *Journal of the American Chemical Society*, **146**, 15070 (2024).
18. S.K. Ha, D. Kalyani, M.S. West, J. Xu, Y.-H. Lam*, et al.*, *Journal of the American Chemical Society*, **147**, 19602 (2025).
19. A.J.S. Hammer, A.I. Leonov, N.L. Bell, and L. Cronin, *JACS Au*, **1**, 1572 (2021).
20. F.H. Loïc M. Roch, Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars P. E. Yunker ,Jason E. Hein, Alán Aspuru-Guzik, *PLOS One*, **15**, e0229862 (2020.
21. R. Mercado, S.M. Kearnes, and C.W. Coley, *Journal of Chemical Information and Modeling*, **63**, 4253 (2023).