Received: 5 July 2023 | Accepted: 6 October 2023

DOI: 10.1002/aaai.12154

SPECIAL TOPIC ARTICLE

Molecule Maker Lab Institute: Accelerating, advancing, and democratizing molecular innovation

Martin D. Burke | **Scott E. Denmark** | **Ying Diao** | **Jiawei Han** | **Rachel Switzky** | **Huimin Zhao**

University of Illinois, Urbana-Champaign, Illinois, Urbana, USA

Correspondence
Huimin Zhao, University of Illinois, Urbana-Champaign, Urbana, IL, USA.
Email: zhao5@illinois.edu

Funding information
National Science Foundation, Grant/Award Number: 2019897

Abstract

Many of the greatest challenges facing society today likely have molecular solutions that await discovery. However, the process of identifying and manufacturing such molecules has remained slow and highly specialist dependent. Interfacing the fields of artificial intelligence (AI) and synthetic organic chemistry has the potential to powerfully address both limitations. The Molecule Maker Lab Institute (MMLI) brings together a team of chemists, engineers, and AI-experts from the University of Illinois Urbana-Champaign (UIUC), Pennsylvania State University, and the Rochester Institute of Technology, with the goal of accelerating the discovery, synthesis and manufacture of complex organic molecules. Advanced AI and machine learning (ML) methods are deployed in four key thrusts: (1) AI-enabled synthesis planning, (2) AI-enabled catalyst development, (3) AI-enabled molecule manufacturing, and (4) AI-enabled molecule discovery. The MMLI's new AI-enabled synthesis platform integrates chemical and enzymatic catalysis with literature mining and ML to predict the best way to make new molecules with desirable biological and material properties. The MMLI is transforming chemical synthesis and generating use-inspired AI advances. Simultaneously, the MMLI is also acting as a training ground for the next generation of scientists with combined expertise in chemistry and AI. Outreach efforts aimed toward high school students and the public are being used to show how AI-enabled tools can help to make chemical synthesis accessible to nonexperts.

INTRODUCTION

The long-term strategic goal of the Molecule Maker Lab Institute (MMLI) is to accelerate, advance, and democratize molecular innovation. To achieve this, the MMLI is creating an open, exciting, and dynamic interdisciplinary ecosystem that will catalyze highly impactful and inclusive collaborations between world-leading PIs, top-notch students, postdocs, and fellows in AI and chemistry, and passionate and creative leaders in education and community engagement. The MMLI is a first-of-its-kind research infrastructure that is having a powerful impact on the

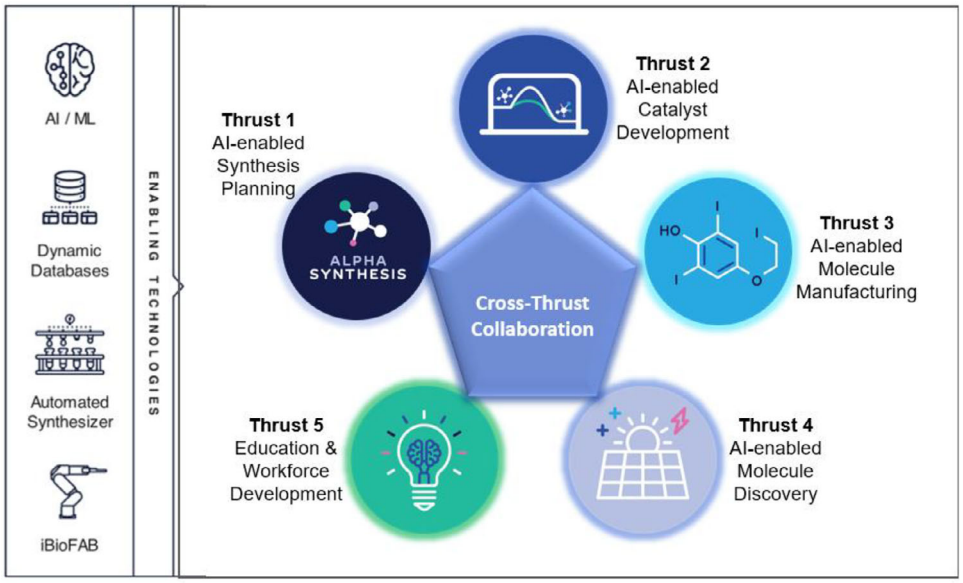

FIGURE 1 Overview of the Molecule Maker Lab Institute.

U.S. research community. The MMLI is achieving a broad impact by tailoring opportunities to a variety of audiences, responding to high priority needs of communities seeking to (1) discover and optimize a wide range of molecular functions, (2) harness the power of data to advance the science of molecular synthesis, and (3) inspire a broad audience of scientists, teachers, students, and citizen scientists to participate in the process of molecular innovation.

USING AI TO ACCELERATE, ADVANCE, AND DEMOCRATIZE MOLECULAR INNOVATION

Many of the greatest challenges facing society today likely have molecular solutions that await discovery. However, the process of identifying and manufacturing such molecules has remained slow and highly specialist dependent. Interfacing the fields of AI and synthetic organic chemistry at the MMLI has the potential to powerfully address both limitations. Through this interdisciplinary initiative, leaders in AI and organic synthesis (both chemical and biological) are intensively collaborating to create frontier AI tools, dynamic open access databases, and fast and broadly accessible small-molecule manufacturing and discovery platforms (Figure 1). Specifically, advanced AI and machine learning (ML) methods are being developed and deployed in the context of four key thrusts focused on: the design of highly effective and, in many cases, Lego-like modular and automatable, synthetic routes for manufacturing and discovering a wide range of small molecules (Thrust 1, AI-enabled synthesis planning), the design and development of optimized chemical and biological catalysts for promoting important reactions with broad potential utility in small-molecule synthesis, including iterative and automated building block assembly processes (Thrust 2, AI-enabled catalyst development), the efficient manufacture of three key molecules already known to perform useful functions, including C2'epi amphotericin B (a novel potent and nontoxic antifungal drug candidate), artemisinin (a critical antimalaria drug), and Millad NX 8000 (an environmentally advantageous colorless, odorless thermoplastic clarifier for polypropylene) (Thrust 3, AI-enabled molecule manufacturing), and the discovery of efficient and stable next generation organic photovoltaics (OPVs) via a fully automated closed-loop autonomous discovery platform (Thrust 4, AI-enabled molecule discovery).

A major challenge and a transformative opportunity for science, engineering, and society lies in bringing the power of making molecules to everyone. The MMLI is developing a versatile and flexible synthesis planning tool named AlphaSynthesis (Figure 2), inspired by AlphaGo and AlphaFold. This tool uses AI to design, construct, and optimize the most effective and automatable synthetic routes, exploiting both chemical and biological catalysts to manufacture target molecules and discover new molecules. The Institute has also created an open-access database that includes all the building blocks, reagents, conditions, and yields (products and byproducts) for every coupling reaction that is run in the MMLI. This database will also have the unique feature of being accessible for on-demand content optimization via automated reaction executions in response to cloud-submitted requests from computer scientists. The MMLI is also developing new AI

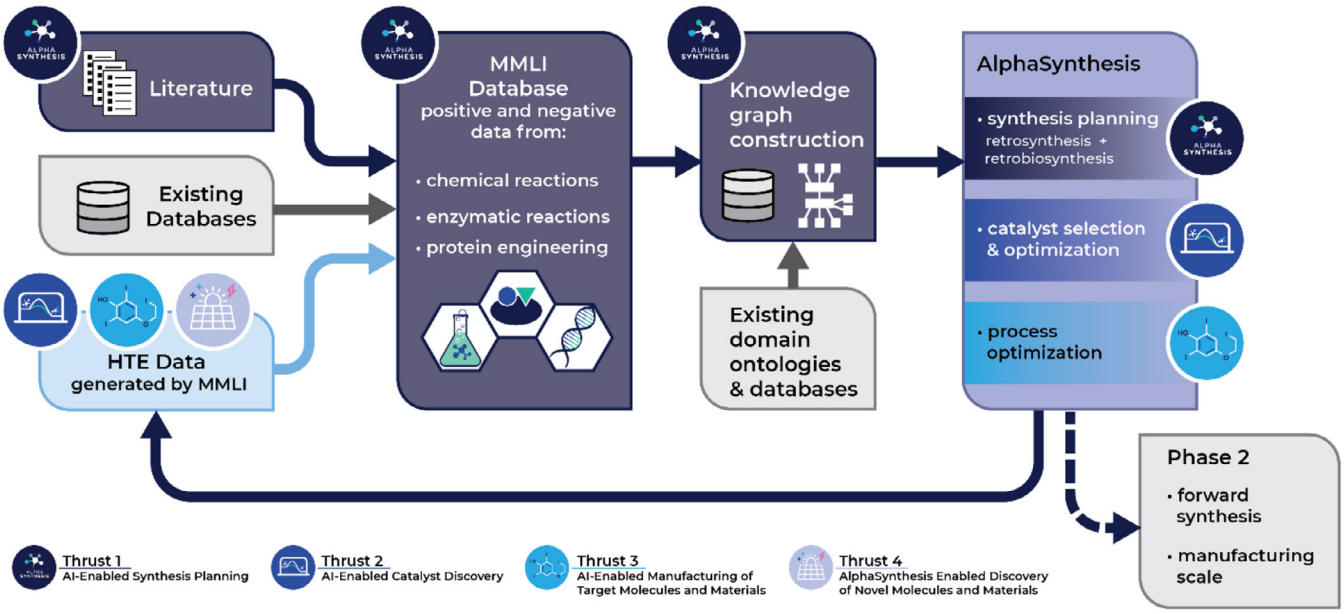

FIGURE 2 Overview of the development of AlphaSynthesis, an artificial intelligence (AI) platform that enables the design, construction, and optimization of the synthetic routes for any molecules.

and ML algorithms and tools for designing and optimizing the catalysts that are required for the implementation of the synthetic routes designed by AlphaSynthesis. To demonstrate the power of these frontier AI tools for synthesis, the MMLI is designing and establishing the most efficient synthetic routes for important molecules and discovering new OPV molecules. These collective activities will increase the efficiency with which small molecules can be manufactured and discovered, drive the advanced development of a wide range of frontier AI methods, and broaden access to the small molecule making process.

Notably, the MMLI is not simply an integrated collection of equipment for automated molecular synthesis and software. It is an open ecosystem of disruptive thinking, education, and community engagement powered by state-of-the art molecular design, synthesis, and spectroscopic characterization technologies all interfaced with a modern cyberinfrastructure.

RECENT RESEARCH ACCOMPLISHMENTS

In Thrust 1, we are focusing on the development of AlphaSynthesis. We are advancing foundational AI research, as highlighted by multiple papers accepted and published by top tier journals such as *Science* and *Nature Catalysis*, and top AI conferences (Angello et al., 2022; Rose et al., 2022; Wang et al., 2022a; Wang et al., 2022b; Yu et al., 2023a; Yu et al., 2023b; Yu et al., 2023c). Beyond publications, MMLI is providing novel foundational AI advances through new information sources and new analytical techniques such as reaction aware multimodal molecular representation and a translator between natural language and molecules. Importantly, this foundational AI work is not confined to the chemistry discipline. These tools and algorithms synergize with and can be applied to other scientific disciplines. Here, we provide three examples to demonstrate this key point.

First, newly developed molecule language models have been built (Edwards et al., 2022) and created results which can be tested in the physical world. Specific focus has been on extracting information from the chemistry literature to enable new insights into how specific structural components of kinase inhibitors, drugs commonly used to treat cancer, are connected to important therapeutic properties, such as blood brain barrier penetration. This work has the potential to yield new AI-based tools to advance the design and discovery of new and improved drugs, which may save researchers years of effort and millions of dollars, not to mention possibly saving lives. Recently, the MMLI has proposed a novel in-context learning framework for personalized drug synergy prediction. This exciting work may eventually enable the creation of a standardized assay for predicting patient-tumor-specific drug synergies, allowing highly targeted combination cancer therapy.

Second, automated extraction of structured knowledge from text-intensive, unstructured scientific literature is a fundamental challenge in AI. To address this challenge, the MMLI is building a ReactIE system which automatically extracts essential and structured chemistry reaction information from chemistry research literature (Zhong

et al., 2023). The structured chemical reaction information includes products, reaction types, reactants, catalysts, solvents, temperatures, and yields, without the assistance of human annotation. The method can be applied to the extraction of other structured information from general scientific literature. The method combines two weakly supervised approaches for pretraining. It utilizes frequent patterns within the text as linguistic cues to identify specific characteristics of chemical reactions. Moreover, it utilizes synthetic data from patent records as distant supervision to incorporate domain knowledge into the model. Experiments demonstrate that ReactIE achieves substantial improvements and outperforms all existing state-of-the-art methods.

Third, the MMLI has created ChemScraper, a formula parser for molecules in born-digital PDF papers, where graphics are defined using PDF drawing instructions (e.g., where characters and lines are given explicitly). Advantages of parsing from born-digital diagrams include avoiding OCR and thereby preventing recognition errors and increasing speed (currently 200 ms/molecule using unoptimized python). The first version of the parser for born-digital molecule diagrams has been completed and ChemScraper has been integrated into the AlphaSynthesis platform.

In Thrust 2, advanced AI and ML methods are being developed and deployed in the design and development of optimized chemical and biological catalysts for promoting important reactions with broad potential utility in small-molecule synthesis. For example, we recently developed an AI tool named Contrastive Learning enabled Enzyme ANnotation (CLEAN) for predicting enzyme function from its sequence (Yu et al., 2023b), which can help the selection of proper enzymes in the designed synthetic routes. Enzyme function annotation is an urgent need and challenge. Many computation tools have been developed including alignment-based methods and ML-based methods (Altschul et al., 1997; Sanderson et al., 2023). However, most of the tools cannot reliably annotate protein functions such as enzyme commission (EC) number, partly due to the fundamental challenge of highly imbalanced enzyme dataset. By contrast, CLEAN can assign EC numbers to query enzyme sequences with better accuracy, reliability, and sensitivity than existing methods as shown in various in silico benchmark studies. Most importantly, we demonstrate that using contrastive learning frameworks, CLEAN can better handle the biased dataset compared to other ML models.

In addition, we have also tested CLEAN on in-house curated halogenase dataset. In particular, we experimentally evaluated three of the halogenases (MJ1651, TTHA0338, and SsFIA) which CLEAN predicted differently with the current annotation in the database. The three halogenase represented three different cases for mis-annotated, un-annotated and partially annotated in the database. In vitro experiments result shows CLEAN has successfully annotated these three halogenases which other methods cannot.

Furthermore, we have implemented a free-to-use web interface version of CLEAN for the broad research community. As part of the MMLI AlphaSynthesis platform, the web tool of CLEAN can be accessed at Clean (https://clean.frontend.mmli1.ncsa.illinois.edu/configuration). After the tool went online, it gained wide community attention. In the first 2 months alone, the web page attracted 20,100 page views and 12,531 of them were from unique viewers. Since the initial release of the website, we have been constantly updating many features, including better result filtering and displaying, computing efficiency, confidence level display, and general user experience. It is worth noting that many of the new features were suggested by our users through the feedback function we provide.

Thrust 3 focuses on leveraging and advancing AI in the context of developing processes for synthesizing on large-scale target molecules with known functions. These include chemical, enzymatic, and hybrid chemoenzymatic routes designed to maximize efficiency in each case. There is specific focus on the development of scalable syntheses for three target molecules; Millad NX8000, C2'epiAmB, and artemisinin, all three of which have important functions driving the need for large-scale access, but none of which have optimal large-scale syntheses. Millad NX8000 is a plastics clarifying agent synthesized yearly on ~6000 metric tons/year scale but requires superstoichiometric tin and generates ~9000 metric tons/year of toxic tin byproducts. C2'epiAmB is an exceptionally promising new renal sparing polyene antifungal, a variation of which has entered clinical trials in the U.S. and there is substantial room for improvement in its current synthesis. Artemisinin is a critically important antimalarial compound on the WHO's list of essential medicines, but all current routes for its manufacture are too expensive to meet global supply demands.

The focus of Thrust 4 is on progress to discover novel OPVs, which can transform the way we harness and use sunlight. The key goal we are working toward is OPVs with high efficiency and high stability. To date, no one has been able to achieve both parameters. The MMLI's approach to address this problem is three-fold. First, the small amount of literature on AI-guided optimization of OPVs focuses on power conversion efficiency. However, the major challenge in this area has become device stability now that the best cell is approaching 20% efficiency. There is a critical gap of relating molecular design to stability which is exactly what the MMLI is addressing. Second, the majority of Al for OPV literature mines published literature or

computed properties, given scarce experimental data that are consistent. Unfortunately, this severely limits the predicting power of AI. Thrust 4 is addressing this issue head on with high throughput syntheses and characterizations to rely primarily on experimentally generated data. Finally, other approaches to OPV discovery use AI optimization of device properties as a function of formulation and processing conditions, but this approach is intrinsically limited to the maximum possible efficiency of the active layer chemistry. The MMLI's innovative strategy uses AI to guide a search through a large synthetic space of OPV chemistries.

The MMLI is targeting OPVs with >10% efficiency and >10 years lifetime to be commercially viable for next-generation energy capture applications and for mitigating climate change. To achieve this goal, autonomous synthesis, automated characterization, and AI-based methods are integrated into a closed-loop approach to drive molecular discovery guided by target criteria for OPV performance: efficiency and stability.

While these thrusts are listed separately it should be noted that there is active and intentional collaboration between thrusts to further enable research and discovery. These collaborations not only increase the impact of the institute but also provide students and postdocs an opportunity for learning the value of collaborating with other research teams.

ENGAGING INDUSTRY PARTNERS

The MMLI has an active Industrial Partnership Program with the goal of providing the opportunity for the two-way exchange of information between the MMLI and industry researchers. The program is a way for MMLI researchers to share the tools and databases being developed for more efficient synthesis and discovery of chemical and materials for a wide-range of applications. Industry researchers can provide perspective on which projects will most benefit society. There is a two-tier membership structure (Partner and Associate), with the annual membership fees going into a seed grant fund to support proof-of-concept proposals from the broader community and expand partnerships. The MMLI continues to grow this program with significant input from the current industry partners.

CULTIVATING MOLECULAR INNOVATORS

As MMLI increases its access and exposure to learners and the general public, it is critical that all of the Thrust 5 initiatives share a common framework with which to situate particular contributions to the larger MMLI identity of democratization. The dimension of identity spans from "Chemical Learners" to "Molecular Innovators". At one extreme of the dimension, chemical learners are focused on a structural and formal exploration of chemistry via MMLI's tools and research. At the opposite extreme, molecular innovators are leveraging the work of MMLI as problem-solving tools focused on the function of molecules and do not engage with formal chemical representations. This span of interaction identities is then framed by the mediums in which they engage with MMLI across the physical and digital resources (in-class labs, camp activities, online lessons, etc.). The goal with this framework is to ensure that MMLI grows not only considering the extremes of the framework, but also to facilitate blending and transitioning between dimensions to provide a more holistic engagement experience whenever possible.

The MMLI is revolutionizing the way chemistry is taught and capture the imagination of a new generation of molecule makers by building on our already established momentum of engagement with educators and students through several mechanisms, including but are not limited to the MMLI in a Box, a Digital Molecule Maker (DMM), as well as establishing international and industrial partnerships. The MMLI aims to democratize the molecule making process and support the next generation of molecule makers, thereby having a wide range of broader impacts.

The MMLI in a Box transforms all the powers of a real lab into carefully crafted low-tech material (e.g., lenticular printed cards, 3D printed macroscopic models), and highly intuitive, engaging hands-on and role-play activities inspired by the latest Next Generation Science Standards guidelines (NGSS Lead States, 2013). These materials and activities empower teachers to help their students enter the awe-inspiring world of molecules, AI, and AI-powered molecule making.

The DMM is the product of a collaboration between the MMLI and the Siebel Center for Design at the UIUC. The DMM interface allows researchers to make their chemical building-blocks available for exploration by learners, who can then combine them while getting dynamic feedback properties of a molecule during its construction. The DMM offers the opportunity for learners to contribute to actual research initiatives by facilitating the connection between a molecular block set, a proposed molecule, and its creation via automated synthesis. The current version of the DMM allows learners to engage with a "10×10×10" (ten start, middle, and end blocks) molecular block set from the Burke Lab at the UIUC. This block set was created to explore molecular structure and light absorption. Already, student molecule submissions are assisting the lab in improving the quality of AI-assisted molecular property predictions and testing the automated synthesis workflow.

The DMM coupled with the MMLI in a Box further supports the development of such intuitions, as well as creativity and imagination, by allowing students to explore thousands of instances of Lego-like building of molecules as they design their own new molecule with novel functions. Finally, synthesizing student-designed molecules allows students to reflect on their intuitions.

MMLI IN THE CLASSROOM

MMLI has partnered with the School of Chemical Sciences at UIUC to implement a three-part sequence into the undergraduate labs. The curriculum in this sequence aims to integrate both data science and automated synthesis into both general and organic chemistry courses. Part 1 is an adaptation of the K-12 version of the MMLI in a Box where students are exposed to the power of molecules, synthesize dyes in a mix-and-match fashion, participate in a novel role-playing game, and combine all that knowledge to analyze a data set from an original research project. During the 2022–2023 Academic Year, ≈1050 students in General Chemistry I and II participated in the MMLI in a Box activity.

In Spring 2023, the second part of the three-part sequence was piloted at the General Chemistry II and Organic Chemistry I level with an activity focused on automating synthesis. Over 100 students came to the Molecule Maker Lab space to participate in the activity to leverage skills that were directly applicable to the work of MMLI, providing a source of inspiration and tangible connection to the science of automated small-molecule synthesis.

CONCLUSION

The MMLI is committed to accelerating, advancing, and democratizing molecular innovation by creating an open, exciting, and dynamic interdisciplinary ecosystem that will catalyze highly impactful and inclusive collaborations between world-leading PIs, top-notch students, postdocs, and fellows in AI and chemistry, and passionate and creative leaders in education and community engagement. The MMLI is transforming chemical synthesis and generating use-inspired AI advances. Simultaneously, the MMLI is also acting as a training ground for the next generation of scientists with combined expertise in chemistry and AI. The long-term impact of the MMLI will be substantial and is expected to shift the paradigm of how molecules are discovered and used to address society's grand challenges. These collective activities will powerfully enable the more efficient manufacturing and discovery of molecules with important functions, drive the advanced development of a wide range of frontier AI methods, and broaden access to the small molecule making process. More information can be found on the MMLI website (https://moleculemaker.org/).

ACKNOWLEDGMENTS

The material is based upon work supported by the National Science Foundation under Grant No. 2019897. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

ORCID

Martin D. Burke https://orcid.org/0000-0001-7963-7140
Scott E. Denmark https://orcid.org/0000-0002-1099-9765
Ying Diao https://orcid.org/0000-0002-8984-0051
Jiawei Han https://orcid.org/0000-0002-3629-2696
Rachel Switzky https://orcid.org/0009-0004-6225-5989
Huimin Zhao https://orcid.org/0000-0002-9069-6739

REFERENCES

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25: 3389–402.

Angello, N. H., C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski, and .D. Burke. 2022. "Closed-loop Optimization of General Reaction Conditions for Heteroaryl Suzuki-Miyaura Coupling." *Science* 378: 399–405.

Edwards, C., T. M. Lai, K. Ros, G. Honke, K. Cho, and H. Ji. 2022. "Translation between Molecules and Natural Language." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP2022).

NGSS Lead States. 2013. *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Rose, T., J. C. Timmerman, S. A. Bawel, S. Chin, H. Zhang, and S. E. Denmark. 2022."High-Level Data Fusion Enables the Chemoinformatically Guided Discovery of Chiral Disulfonimide Catalysts for Atropselective Iodination of 2 Amino-6-arylpyridines." *Journal of the American Chemical Society* 144: 22950–64.

Sanderson, T., M. L. Bileschi, D. Belanger, and L. J. Colwell. 2023. "ProteInfer: Deep Networks for Protein Functional Inference." *eLife* 12: e80942.

Wang, H., W. Li, X. Jin, K. Cho, H. Ji, J. Han, and M. Burke. 2022b. "Chemical-Reaction-Aware Molecule Representation Learning." In Proceedings of the International Conference on Learning Representations (ICLR2022).

Wang, X., V. Hu, M. Jiang, Y. Zhang, J. Xiao, D. Loving, H. Ji, M. Burke, and J. Han. 2022a. "ReactClass: Cross-Modal Supervision for Subword-Guided Reactant Entity Classification." In

Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'22), Las Vegas, NV.

Yu, T., A. Boob, N. Singh, Y. Su, and H. Zhao. 2023c. "In vitro Continuous Protein Evolution Empowered by Machine Learning and Automation." *Cell Systems* 14: 633–44.

Yu, T., A. G. Boob, M. J. Volk, X. Liu, H. Cui, and H. Zhao. 2023a. "Machine Learning-Enabled Retrobiosynthesis of Molecules." *Nature Catalysis* 6: 137–51.

Yu, T., H. Cui, J. Li, Y. Luo, G. Jiang, and H. Zhao. 2023b. "Enzyme Function Prediction Using Contrastive Learning." *Science* 379: 1358–63.

Zhong, M., S. Ouyang, M. Jiang, V. Hu, Y. Jiao, X. Wang, and J. Han 2023. "ReactIE: Enhancing Chemical Reaction Extraction with Weak Supervision." In Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics (ACL'23), Toronto, Canada.

How to cite this article: Burke, M. D., S. E. Denmark, Y. Diao, J. Han, R. Switzky, and H. Zhao. 2024. "Molecule Maker Lab Institute: Accelerating, advancing, and democratizing molecular innovation." *AI Magazine* 45: 117–23. https://doi.org/10.1002/aaai.12154

AUTHOR BIOGRAPHIES

Martin D. Burke is a Professor of Chemistry at the UIUC and leads MMLI Thrust 3.

Scott E. Denmark is a Professor of Chemistry at the UIUC and leads MMLI Thrust 2.

Ying Diao is an Associate Professor of Chemical and Biomolecular Engineering at the UIUC and leads MMLI Thrust 4.

Jiawei Han is a professor of Computer Science at the UIUC and leads MMLI Thrust 1.

Rachel Switzky is the inaugural director of the Siebel Center for Design at the UIUC and leads MMLI Thrust 5.

Huimin Zhao is a professor of chemistry, biochemistry, biophysics, and bioengineering at the UIUC and is the director of the MMLI.
For more information on the authors, see the MMLI website (moleculemaker.org).