Integrating Machine Learning and Large Language Models to Advance Exploration of Electrochemical Reactions

Zhiling Zheng,[†] Federico Florit,[†] Brooke Jin,[†] Haoyang Wu,[†] Shih-Cheng Li,[†] Kakasaheb Y. Nandiwale,[§] Chase A. Salazar,[§] Jason G. Mustakis,[§] William H. Green,[†] and Klavs F. Jensen[†,*]

[†] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142, United States

[§] Chemical Research & Development, Pfizer Worldwide Research and Development, Groton, CT 06340, United States

KEYWORDS: chemical reactions, machine learning, large language models, hydrocarbons, oxidation.

ABSTRACT: Electrochemical C-H oxidation reactions offer a sustainable route to functionalize hydrocarbons, yet the identification of competent substrates and their synthesis optimization remains challenging. Here, we report an integrated approach combining machine learning (ML) and large language models (LLMs) to streamline the exploration of electrochemical C-H oxidation reactions. Utilizing a batch rapid screening electrochemical platform, we evaluated a wide range of reactions, initially classifying substrates by their reactivity, while LLMs text-mined literature data to augment the training set. The resulting ML models, one for reactivity prediction and the other one for site selectivity, both achieved high accuracy (>90%) and enabled virtual screening of a large set of commercially available molecules. To optimize reaction conditions of substrates of interest upon the screening, LLMs were prompted to generate code to iteratively improve yield, lowering the barrier for scientists to access ML programs, and this strategy efficiently identified high-yield conditions for eight drug-like substances or intermediates. Notably, we benchmarked the accuracy and reliability of 10 different LLMs, including llama, Claude, and GPT-4, on generating and executing codes related to ML based on natural language prompts given by chemists to showcase their tool-making and tool-using capabilities and potentials for accelerating research across four diverse tasks. In addition, we collected an experimental benchmark dataset comprising 1071 reaction conditions and yields for electrochemical C-H oxidation reactions, and our findings revealed that integrating LLMs and ML outperformed using either method alone. We envision that this combined approach offers a robust and generalizable pathway for advancing synthetic chemistry research.

INTRODUCTION

Electrochemical C-H oxidations are tunable and cost-effective transformations for streamlining conversion of hydrocarbons to decorated oxidized molecules.[1–4] As synthetic chemists actively expand the scope of this field and discover new chemical reactions, the selection of reactive substrates and the subsequent optimization of synthesis parameters, while often guided by fundamental chemical principles and hypotheses, still require extensive empirical condition screening and remain resource-consuming. Consequently, smart workflows that bypass the traditional trial-and-error approach are essential to meet the increasing demand from chemists to navigate reactivity space and expedite new reaction discovery.[5]

Recent advancements in machine learning (ML) have shown promising potential in reactivity prediction and optimization of organic reactions.[6–10] Simultaneously, large language models (LLMs) have gained attention in chemical research, offering novel interfaces with chemical data and assisting chemists—particularly those less experienced with digital tools—in customizing their digital workflows through intuitive natural language prompts.[11–15] In essence, ML provides a toolbox for specific tasks in chemical research and LLMs serve as meta-tools that enhance the accessibility of these computational tools, bridging the gap between digital proficiency and chemical expertise. Consequently, the integration between the mathematical rigor of ML and the language understanding and domain-specific knowledge of LLMs can make data-driven methodologies accessible to a broader community of chemists. This motivates us to explore the following underexplored questions in this evolving field: What are the capabilities of LLMs in synthetic electrochemistry? How can LLM be integrated with ML to reliably expedite reaction discovery?

Herein, we demonstrate the synergistic potential of ML and LLMs to advance the exploration and optimization of electrochemical C-H oxidation reactions. For the ML aspect, our approach aims to address two fundamental questions: (1) Which substrates are suitable for electrochemical oxidation? and (2) What synthesis conditions give optimal results? By leveraging both literature data and rapid screening experimental results, we train models to predict reactivity and selectivity for C(sp^3)–H oxygenation, enabling *in silico* screening of chemical entities for initial hits. Subsequently, for these selected substrates of interest, an active learning protocol designed for reaction yield optimization was applied to iteratively and rapidly navigate the search space to identify optimal synthesis conditions. Throughout this study, we also illustrate the versatility of LLM agents—to be a tool (e.g. extracting knowledge from literature), to use tools (e.g. employing ML and liquid handler for synthesis), and to create tools (e.g. generating custom Python code for prediction and analysis). As such, this multifaceted utility—enhancing workflow efficiency and intelligence—serves as a practical example of the application of LLMs in streamlining chemical research processes and underscores their potential to accelerate scientific discovery for synthetic chemists.

RESULTS AND DISCUSSION

Rapid Screening Electrochemical Platform

At the onset of this study, the goal was to assess whether ML could guide the selection of compounds suitable for electrochemical C(sp^3)-H oxidation. To accomplish this, we required a diverse training set composed of both substrates amenable to electrochemical oxidation and those that are not—modeling requires both positive and negative training data sets for better accuracy in prediction. Most reactions available in the literature predominantly include successful examples of reactive and high-yield substrates, leaving a gap in data on unsuccessful conditions, which are equally critical for training robust predictive models.

To address this issue, inspired by previous works on electrochemical synthesis platforms,[16–18] we developed a rapid screening electrochemical platform capable of conducting multiple reactions simultaneously, thus enabling both reactivity screening and optimization of synthesis conditions. This platform features a standardized 24-well plate electro-synthesis reactor (Figure 1A), which includes a water jet-cut anode and cathode connectors, an alignment plate, and a vial locator. Components for the reactor are readily available through commercial vendors and can be easily assembled at low cost in the lab (See Supporting Information Section S2 for detailed design and assembly information). Furthermore, the choice of electrodes can be adjusted to meet specific experimental needs, enhancing the flexibility of our setup. The experimental setup includes 4 mL electrochemical reactors with two pairs of counter electrodes, totaling four electrodes (Figure 1A), and it allows for increased surface area and improved current distribution, which enhances mass transport. Such improvements facilitate the diffusion of reactants to the electrode surfaces and the efficient removal of products. Additionally, dual-electrode setup ensures that the reaction can continue even if one pair fails or loses connection, thus improving the reliability of the experimental setup.

Initially, our platform was employed for reactivity screening to acquire labeled data points essential for training our machine learning models about substrate suitability for C(sp^3)-H oxidation. To ensure a diverse chemical space in our training dataset, using a similar approach reported in previous literature[5], we randomly selected 335 chemicals available in our laboratory and subjected them to predetermined electrochemical conditions[19,20] to enable rapid generation of data points, allowing us to classify each substrate based on its reactivity (Section S2.2 of the Supporting Information). In particular, the reactions targeted the transformation of substrates into ketone or alcohol products, using mediator-catalyzed C(sp^3)-H bond oxidation, with the outcomes verified through NMR spectroscopy by monitoring the appearance of signature peaks indicative of these products.

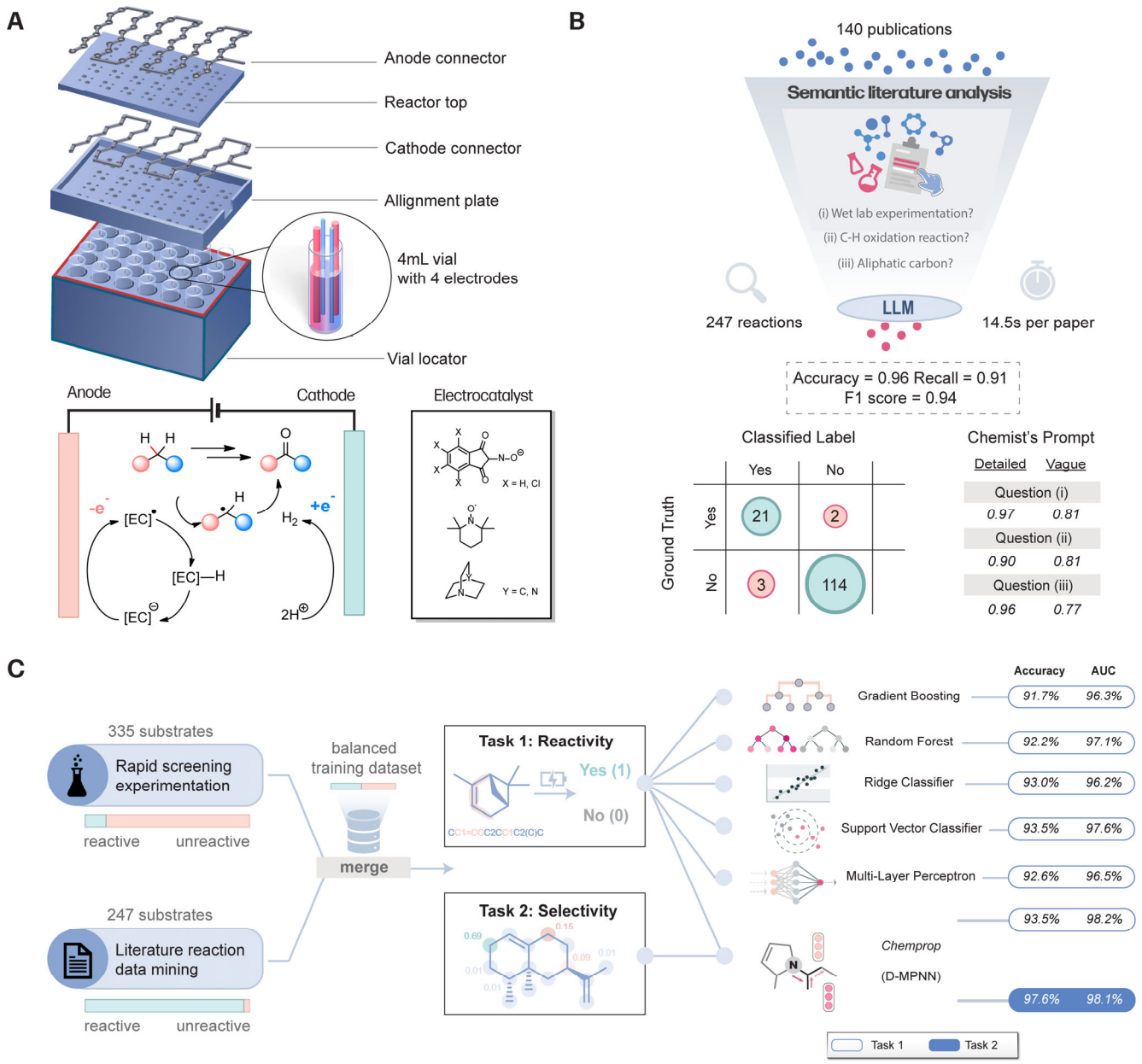

Figure 1. (**A**) Design and assembly of the 24-well electrochemical platform and the schematic overview of the electrochemical C(sp³)-H oxidation process using the electrocatalyst. (**B**) Semantic literature analysis for reaction data mining using a language model with natural language prompts. Performance is evaluated by comparing the ground truth with LLM-assigned labels and examining the impact of prompt quality. (**C**) Overview of training data preparation and machine learning models used for predicting electrochemical C-H oxidation reactivity (Task 1) and selectivity (Task 2). Models with different architectures are evaluated for accuracy and AUC.

As the first step to constructing a balanced and informative dataset, we assigned negative labels to reactions where the transformation either did not occur or resulted in unknown products beyond the scope of our study. This approach allowed us to simplify the modeling challenge to a binary classification task, focusing solely on predicting substrate reactivity under a fixed set of reaction conditions. This strategy avoided the complexities associated with predicting novel products and was sufficient to enable the machine learning model to focus on learning patterns relevant to the specific type of electrochemical transformation in this study (Figure 1A). We acknowledge that while predetermined conditions generally work for a number of known reactants, they could yield low outcomes during screening. However, we designated yield optimization as a separate task later in this study to streamline the initial reactivity screening process and avoid the need to tune numerous parameters of electrochemical reactions during this phase. The deployment of this standardized and low-cost rapid screening electrochemical reactor allowed rapid acquisition of a 335 experimental electrochemical oxidation outcome dataset (Table S2). which includes both positive and negative labels.

Literature Data Mining

In parallel to collecting experimental data, we retrieved reaction data and outcomes from the scientific literature to augment the C-H oxidation reaction dataset with a diverse choice of candidates. Traditional approaches to querying reaction databases often struggle to capture the nuanced criteria specific to our study, such as focusing on electrochemical C-H oxidation reactions on aliphatic carbons using a mediator. For a specific demand on the reaction, manually analyzing and curating a large corpus of papers to extract relevant examples that meet these criteria would typically be very time-consuming for a human.

To address this challenge, we employed semantic analysis using LLM agents guided by human instructions. This approach allows for precise extraction of relevant data by understanding and interpreting the context within scientific manuscripts. Specifically, the LLMs were tasked with identifying papers that met three critical criteria for our electrochemical oxidation dataset: (1) the paper must be an experimental study on electrochemical synthesis (i.e., not a review paper or computational study), (2) it must involve C-H bond oxidation to alcohol or ketone products (i.e., not other types like C-C coupling), and (3) the reaction must occur on an aliphatic carbon (i.e., not a C(sp^2)-H) (Section S3.2 of the Supporting Information). In essence, the LLMs function as a customized filter, guided by human language instructions, that automates the process of reading through each paper, understanding not only the experimental sections but also the discussions, and selecting the qualifying papers. To evaluate the performance, we analyzed 140 relevant papers using pre-designed prompts (Table S4). The LLMs took approximately 15 seconds per manuscript to assign a Yes/No label, resulting in a total analysis time of about 35 minutes for all selected literature (Figure S18). Validation against ground truth revealed that the LLMs achieved 96% accuracy, correctly identifying 21 relevant papers containing 497 reactions for 247 substrates (Figure 1B).

The power of semantic analysis using LLMs is further underscored by their ability to provide reasoning for their decisions (Figures S15 and S16).[11,21–23] For each of the three criteria, the LLM justifies its answers based on specific paragraphs or sentences in the original literature, leveraging the reasoning capabilities of large language models. To ensure minimal hallucination and sound chemical knowledge, we devised specific prompts tailored to the goal of this study and manually analyzed the reasoning statements (Section S3.3 of the Supporting Information), confirming over 90% correctness in decision-making (Figure 1B) and in pointing back to the relevant sections of the original manuscripts. Simultaneously, we conducted an ablation study where shorter, less detailed prompts were used. These prompts, lacking strict instructions to reference the original literature and with less specificity on each question, came with more ambiguity in the instruction and resulted in decreased performance of the LLMs, highlighting the importance of detailed and specific prompts for guiding LLMs, particularly in semantic analysis.

Machine Learning Models Training

Upon completion of dataset collection from experimental outcomes and literature data mining, we amalgamated the data to create a balanced dataset suitable for model training (Figure 1C). The literature data, biased towards successful substrates, complemented the failure data points generated from our screening platform. This combination resulted in a dataset comprising 582 substrates, with 271 oxidizable (46.6%) and 311 non-oxidizable (53.4%) towards electrochemical C-H oxidation reactions (Figure 1C). The dataset included 7,720 carbon atoms, 431 of which were oxidized during the transformations (a complete list is available in the Supporting Information). Our objective was to develop two classes of predictive models: (i) the reactivity prediction model to classify substrates as reactive or non-reactive, and (ii) the selectivity prediction model to classify each carbon atom within a molecule as oxidized or unchanged. The former allows for rapid screening of chemical catalogs, while the latter helps chemists identify which sites are likely to undergo oxidation. After optimizing the respective hyperparameters, each model's performance was rigorously tested using accuracy and Area Under the Curve (AUC) metrics. All models demonstrated high performance, with accuracies over 91.7% and AUC values of 97.2%. These results align well with our previous studies using ML to predict electrochemical reaction outcomes.[5]

Additionally, we explored the inclusion of density functional theory (DFT) descriptors in the Chemprop model (Section S4.1 of the Supporting Information), which provided richer quantum mechanical information and slightly enhanced model performance (Table S10). Besides, Chemprop's interpretability features offered a transparent view of which molecular substructures or features were driving the predictions, guiding the design of new experiments by indicating impactful molecular modifications (Section S4.1 of the Supporting Information). This interpretability advantage led us to select Chemprop for subsequent catalog screenings to make rapid reactivity prediction on over 500,000 commercially available compounds (Figure S21).

Due to these advantages, we chose Chemprop for the selectivity model. For substrates identified as reactive, the selectivity model predicted specific chemical features indicative of selectivity. This task underwent similar performance evaluations, achieving an accuracy of 97.6% and an AUC of 98.1% (Figure 1C and Section S4.2 of the Supporting Information). In this case, the graph neural network's directed-message passing mechanism proved effective in capturing intricate molecular features relevant to selectivity. Collectively, the consistent performance of ML models on the two tasks underscored the robustness of our integrated balanced dataset and the benefit of combining wet lab experimentation with literature data mining.

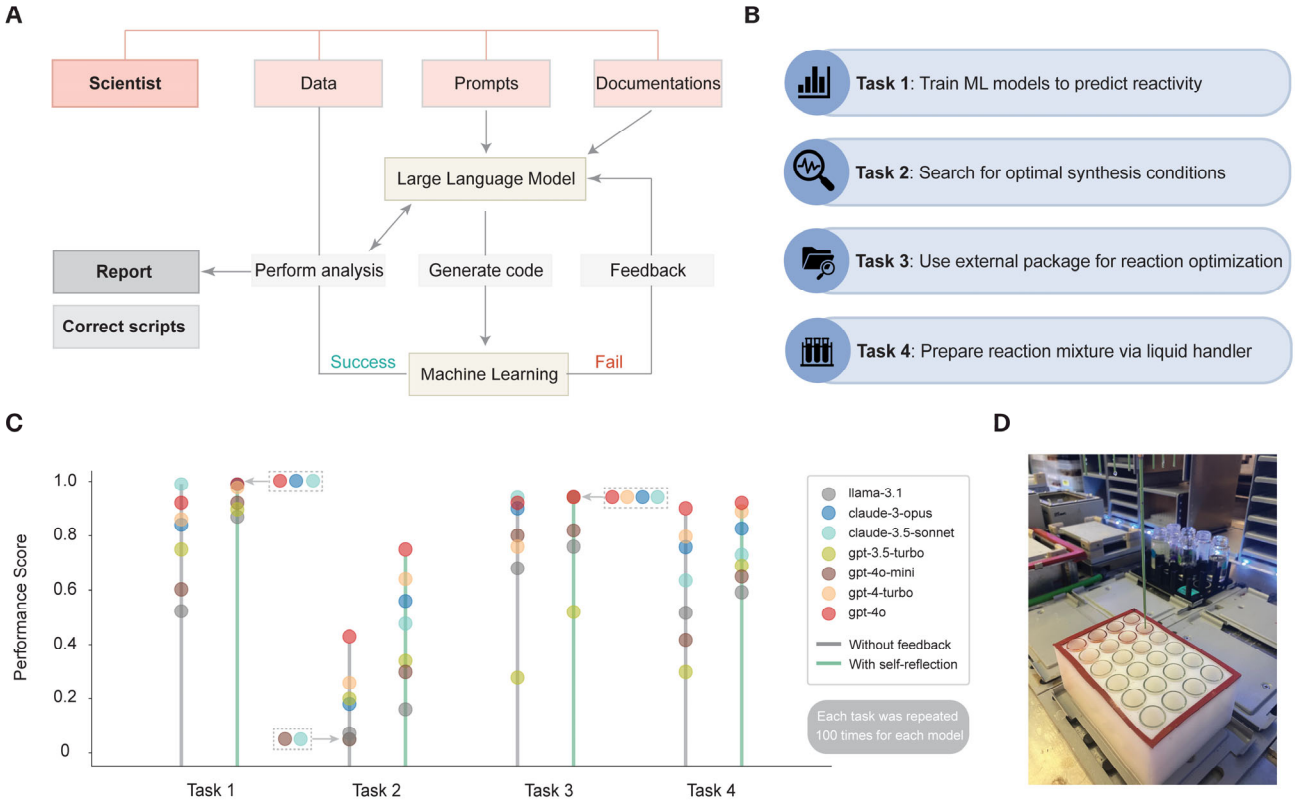

Figure 2. (**A**) Overview of the human-LLM interaction for designing and creating research tools to process data for chemists. (**B**) Prompt-to-code tasks guiding the LLM to develop ML programs or executable code in the context of chemical research. Examples of code generated by LLM for task 1 to 4 are shown in Section S5 of the Supporting Information. (**C**) Comparison of various LLMs on code-writing tasks. Each task was run with single-shot (grey bar) or self-reflection (green bar) approaches. Performance was evaluated by repeating the prompt 100 times and calculating the accuracy. Details can be found in Tables S12–14 of the Supporting Information. (**D**) LLM interprets suggested experimental conditions via ML program and converts them into physical actions for a liquid handler on a robotic platform.

Benchmarking LLMs Auto Code Generation Performance

In the progression of integrating machine learning with chemical research, a pivotal enhancement is the utilization of LLM to automatically generate code for the practical implementation of ML models. Synthetic chemists are often not experts in data science and can have limited coding experience necessary for leveraging machine learning effectively. Consequently, methods that not only understand the chemical context but also automate the coding process needed to process and analyze data could be helpful to the chemistry community. LLMs such as Llama[24], GPT[25], and Claude[26] models offer a promising solution by generating executable code from natural language prompts, potentially lowering the barrier to computational techniques in chemistry and enhancing chemists' productivity in using ML models without extensive coding experience.[11,15,27,28]

Toward this end, as the first step, we developed a "prompt-to-code" framework and used it to evaluate the performance of different open-source and proprietary LLMs in tool-making and tool-using (Figure 2A). The core objective was to assess the reliability and accuracy of code produced by LLMs across four distinct tasks in the context of this study: (1) ML model training using a dataset on C-H oxidation, (2) development of code for tuning synthesis conditions and optimizing reaction yields, (3) interpretation of documentation and application of existing Python package for yield optimization, and (4) direct interaction with laboratory hardware[29] to prepare solutions based on generated synthesis parameters (Figure 2B). These tasks were designed to span a range of practical applications, from data handling to physical lab automation, reflecting the diverse ways LLMs can implement code for ML to support chemical research (Figure S22). Notably, previous evaluations of LLM code-writing performance have often relied on qualitative assessments by human reviewers and have usually been based on one-time conversations, which could introduce bias and did not account for the inherent variability and occasional inaccuracies (hallucinations) in LLM outputs. To address these issues, we developed a rigorous, quantitative benchmarking process using four Python-based code evaluators tailored for each task to not only check the executability of the code but also assess its correctness in a simulated environment (Section S5 of the Supporting Information). To this regard, 10 LLMs were chosen (Full list shown in Tables S12–14 of the Supporting Information) and each LLM was evaluated by repeatedly

generating code for the same task 100 times independently, a robust sample size that mitigates performance variability and reflects the long-term reliability of each model across seven different LLMs.

The results from this benchmarking demonstrated the potential of using LLMs as code assistants to implement ML models for chemists (Figure 2C). For task 1, which involved training ML models, the LLMs demonstrated a high degree of competency (Table S9), with code generation accuracy frequently surpassing 90%. This indicates a strong understanding of the ML frameworks and the ability to apply them correctly to chemical datasets. In task 2, the LLMs faced the more complex challenge of optimizing chemical synthesis conditions (Table S10). Here, the more advanced LLMs (e.g., GPT-4o) showed impressive adaptability, with a success rate in over 60% of the trials, highlighting their potential to handle complex, context-dependent coding tasks. Task 3 tested the LLMs' ability to comprehend and apply unfamiliar Python packages for yield optimization. Here, the LLMs proved to be proficient learners, quickly adapting to new documentation and examples to produce ready-to-use code (Table S11). Finally, in task 4, LLMs were tasked with generating executable scripts for liquid handling robots based on the suggestion made by Taks 2 or Task 3. This task demonstrated the practical applicability of LLM-generated code in automating physical processes in the lab, consistent with the previous literature findings[13,15,30], with successful execution reflecting the LLMs' capacity to integrate digital and physical workflows effectively (Figure 2D).

Furthermore, we introduced a "self-reflection" mode in the benchmarking process and evaluated the codes generated with or without this model (Figure 2A and Figure 2C). When an LLM's generated code failed initial execution, the error message was automatically sent to the LLM, prompting it to modify its output in real-time. It was observed that, for all models, regardless of the size and if it is open-source or not, this iterative process significantly enhanced the quality of the generated code by reducing the hallucinations (Figure 2C), proving that LLMs can learn from their errors and improve subsequently generated codes.

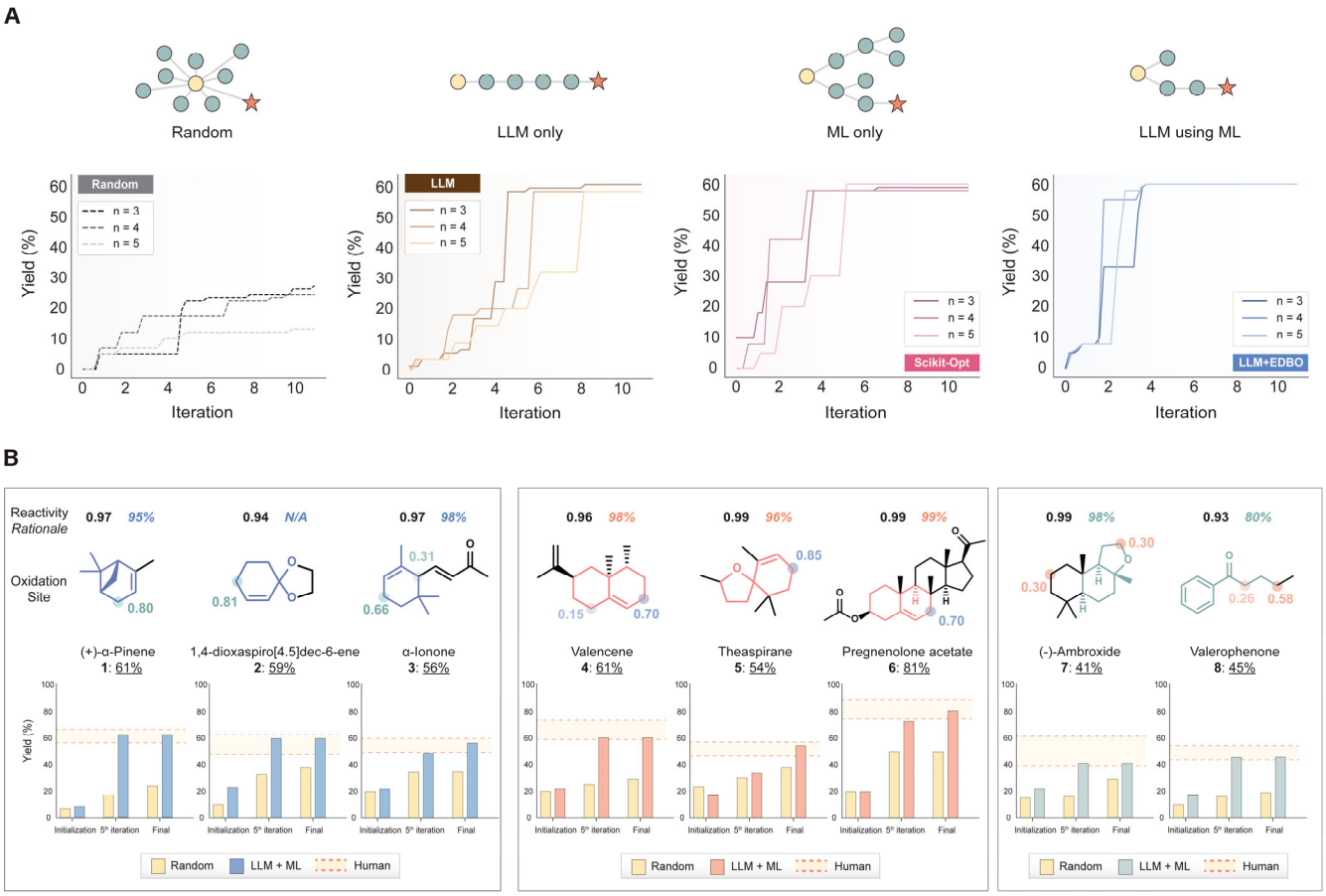

Figure 3. (**A**) Illustration of different approaches in searching for optimal synthesis conditions, along with the performance of each approach on the electrochemical oxidation of α-pinene to verbenone, with "n" indicating the number of reactions per iteration. The combination of LLM and ML leverages both domain knowledge in chemistry and statistical learning. (**B**) Selected substrates based on ML predictions for reactivity and selectivity, along with the resulting optimization process and yields. Substrates are grouped by interpretable substructures and reaction yields. Human-level performance was estimated by reproducing conditions from literature[19,20] or manually optimizing new reactions (Table S28). Each reaction batch size was 4, with a total of 88 reactions per substrate over 10 iterations plus initialization.

Electrochemical Reaction Yield Optimization with Activate Learning Approach

To further our exploration into the synthesis optimization of electrochemical C-H oxidation, we developed and compared several methodologies on the same electrochemical synthesis platform mentioned in the prior section. Our focus shifted towards active learning strategies designed to iteratively refine synthesis conditions to maximize yield while minimizing experimental iterations. We used a batch approach on the screening electrochemical reactor with each batch comprising 3 to 5 reactions, analyzed and adjusted based on the NMR yield outcomes to guide subsequent experimental conditions (Figure 3A). We developed and examined four distinct strategies: (i) random sampling, representing traditional trial-and-error; (ii) LLM-driven prediction[13,23,31], which mimics the decision-making process by human chemist and leverages chemical intuition without statistical learning; (iii) ML-only optimization[8,32], which applies a purely statistical approach devoid of initial chemical insight; and (iv) a hybrid LLM-ML approach, where LLMs guide the initial parameter selection, subsequently LLMs use ML as helper functions to make suggestions on synthesis parameters.

The effectiveness of these methods was first tested on α-pinene due to its high predicted reactivity score (0.87) and selectivity (0.90), making it an ideal candidate for methodological comparison. Over the course of 455 reactions, distributed across 10 iterative rounds per method within the search space (Table 1), we closely monitored the improvement in yield (Figure S39). Our findings indicate that the random method stagnated at low yields (around 20%) even after extensive iteration, underscoring the inefficiency of non-guided experimental approaches. While there is excitement and interest in using LLMs for synthesis optimization, it is important to understand that their "suggestions" are not based on statistical learning and lack a mathematical foundation. Instead, they make educated guesses based on observed general trends and domain knowledge, and usually change one factor at a time (Figures S52 and S53). However, this does not mean that LLMs are not valuable for synthesis optimization; on the contrary, it was demonstrated that the integrated LLM-ML approach can start with an LLM-informed search space that incorporates both literature-derived insights and empirical data. This method rapidly refined the reaction conditions through ML algorithms. In this case, LLM relies on the output from ML models to make decisions rather than having an educated guess on what the next conditions should be tried. This synergy also enabled the precise tuning of conditions to achieve yields over 60% (Figure 3A and Section S6 of the Supporting Information).

Table 1. Optimization synthesis parameters and search space for electrochemical C-H oxidation reactions.

Synthesis Parameter[a]	Choice[b]	Number
Concentration (mM)	25, 50, 75, 100, 125	5
Electrocatalyst	NHPI, TCNHPI, Quinuclidine, DABCO, TEMPO	5
Equivalence of Electrocatalyst	0, 0.25, 0.5, 0.75, 1	5
Electrolyte	$LiClO_4$, LiOTf , Bu_4NClO_4, Et_4NBF_4, Bu_4NPF_6	5
Solvent	ACN, ACN/HFIP (19:1)	2

[a]The reactions were carried out using graphite or RVC anode and nickel cathode, with a potential of 3.5V, at room temperature for 12 hours. The reaction volume was 4 ml, with stirring at 600 rpm. Detailed procedures are provided in Section S6 of the Supplementary Information. [b]Abbreviations: NHPI = N-hydroxyphthalimide, TCNHPI = tetrachlorophthalimide, DABCO = 1,4-diazabicyclo[2.2.2]octane, TEMPO = 2,2,6,6-tetramethylpiperidinyloxy, ACN = acetonitrile, HFIP = hexafluoroisopropanol.

Building on the success with α-pinene, we applied the LLM-ML framework to optimize the synthesis conditions for eight additional substrates (Table S13) with a batch size of four for 10 iterations to demonstrate its generalizability. Notably, eight substrates were identified using the reactivity and selectivity models (Figure 3C), which provide chemists with not only a numerical value indicating the likelihood of oxidation under C-H oxidation but also visualizations of the substructures contributing to oxidation, making the decision-making process more transparent. The selectivity model also indicates which sites are likely to be oxidizable.

In total, we successfully identified the best combination of electrocatalyst, electrolyte, and concentration for reaction (Figure 3C) from 1,250 possible combinations within 10 iterations for each of the 8 chemicals, amounting to a 10,000-reaction space. We note that each compound was independently optimized from the same search space (Table 1). The optimization results demonstrated that all substrates achieved yields comparable to those obtained through human-level optimization, without initial input from a chemist. Interestingly, while some optimal conditions mirrored those reported in the literature, others revealed new, efficient combinations not driven by traditional chemical intuition (Tables S15—28). This balance

between exploitation (fine-tuning in specific regions) and exploration (testing novel combinations) underscored the robustness of the LLM-ML approach.

Additionally, we found that the optimized conditions were often specific to each substrate, as demonstrated by cross-application of synthesis conditions (Figure 4). Each substrate's highest yield was achieved under its uniquely optimized parameters, highlighting the necessity of a tailored approach rather than a one-size-fits-all methodology. This specificity is crucial, as conditions optimized for one substrate are different and did not necessarily translate to others for the optimal reaction yields (Table S14). By dynamically adapting and optimizing reaction conditions, our approach reduces the trial-and-error inherent in traditional methods, enhancing efficiency and productivity in chemical synthesis. The active learning framework proved effective in streamlining the optimization process, effectively reducing the experimental burden while achieving high yields with the active learning approach.

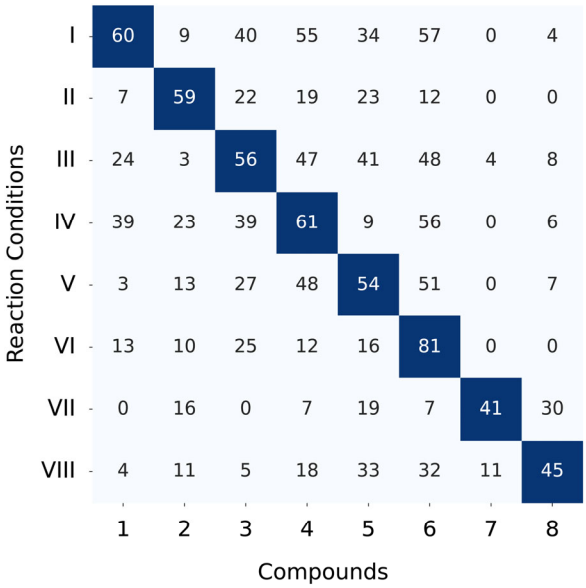

Figure 4. Impact of condition-specific optimization on yield outcomes for electrochemical C-H oxidation reactions. The heatmap illustrates the observed reaction yields (%) of eight different compounds (1 to 8) under various reaction conditions (I to VIII). The deep blue diagonal cells indicate the yields achieved using the unique optimized conditions for each specific compound, determined through the active learning approach. The off-diagonal light blue cells show the yields when optimized conditions for one compound are applied to others. Detailed reaction conditions are available in Supporting Information Table S17, and the search space parameters are listed in Table 1.

CONCLUSIONS

We have successfully (1) developed and validated machine learning models for predicting reactivity and site selectivity in electrochemical C-H oxidation reactions, achieving high accuracy, (2) created a cost-effective, rapid screening electrochemical platform to facilitate rapid data generation and reactivity screening, (3) leveraged large language models to semantically analyze scientific literature and generate ML code, significantly lowering the barrier for chemists to utilize ML tools, and (4) employed a synergistic approach combining ML and LLMs to iteratively refine synthesis conditions, leading to high-yield optimizations for selected substrates and a 1071 electrochemical reaction dataset. At a fundamental level, large language models can be perceived as motivated learners, while their foundational versions might only grasp the basics of chemistry. Importantly, incorporating machine learning models and coherent human instruction significantly enhances their proficiency in a range of chemistry-related tasks. This integration has shown potential in streamlining laboratory work and making digital tools more accessible to chemists with limited coding experience. Despite these promising results, there is still a long journey ahead. Opportunities for improvement include designing more sophisticated human-AI interaction frameworks, better interfacing with other digital tools, and expanding the knowledge base with external sources. Besides, fine-tuning LLMs to better align with human instructions can reduce uncertainty and enhance goal alignment. This integrated AI-powered methodology not only streamlines the traditional trial-and-error process but also offers a robust and generalizable pathway for potentially expanding to a wider range of reaction types and conditions and enable further automating the discovery process by coupling with commercial molecule databases. Overall, this work underscores the potential of human-AI collaboration, combining the strengths of LLMs and ML, to advance synthetic chemistry research.

ASSOCIATED CONTENT

Supporting Information. General experimental, characterization data, spectra, and computational methods; Datasets; Example python codes.

AUTHOR INFORMATION

Corresponding Author

Klavs F. Jensen – *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States; orcid.org/0000-0001-7192-580X; Email: kfjensen@mit.edu*

Other Authors

Zhiling Zheng – *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States; orcid.org/0000-0001-6090-2258*

Federico Florit – *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States; orcid.org/0000-0002-6484-4953*

Brooke Jin – *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States; orcid.org/0009-0009-0779-0133*

Haoyang Wu – *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States; orcid.org/0000-0002-0644-7554*

Shih-Cheng Li – *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States; orcid.org/0000-0001-8645-0034*

Kakasaheb Y. Nandiwale – *Chemical Research & Development, Pfizer Worldwide Research and Development, Groton, Connecticut 06340, United States; orcid.org/0000-0002-0754-7362*

Chase A. Salazar – *Chemical Research & Development, Pfizer Worldwide Research and Development, Groton, Connecticut 06340, United States; orcid.org/0000-0002-2221-4865*

Jason G. Mustakis – *Chemical Research & Development, Pfizer Worldwide Research and Development, Groton, Connecticut 06340, United States; orcid.org/0000-0002-8683-0691*

William H. Green – *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States; orcid.org/0000-0003-2603-9694*

Author Contributions

The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

There are no conflicts to declare.

ACKNOWLEDGMENTS

This material is based upon work supported by Pfizer. The authors extend their gratitude to the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium for their support. Z.Z. is grateful to the OpenAI Researcher Access Program for subsidized access. The authors thank Drs. Andrew Zahrt, Jakob Dahl, Seung Kyun Ha, and Mr. Leo Maeser (Jensen Research Group) for their valuable discussions. Z.Z. expresses gratitude to Drs. Brent Koscher and Matthew McDonald (Jensen Research Group) for their assistance in setting up reactions on the autonomous chemical discovery platform, and to Wenhao Gao (Coley Research Group) for helpful discussions on synthesis optimization.

REFERENCES

(1) Zhu, C.; Ang, N. W. J.; Meyer, T. H.; Qiu, Y.; Ackermann, L. Organic Electrochemistry: Molecular Syntheses with Potential. *ACS Cent. Sci.* **2021**, *7* (3), 415–431. https://doi.org/10.1021/acscentsci.0c01532.

(2) Kingston, C.; Palkowitz, M. D.; Takahira, Y.; Vantourout, J. C.; Peters, B. K.; Kawamata, Y.; Baran, P. S. A Survival Guide for the "Electro-Curious." *Acc. Chem. Res.* **2020**, *53* (1), 72–83. https://doi.org/10.1021/acs.accounts.9b00539.

(3) Yan, M.; Kawamata, Y.; Baran, P. S. Synthetic Organic Electrochemical Methods Since 2000: On the Verge of a Renaissance. *Chem. Rev.* **2017**, *117* (21), 13230–13319. https://doi.org/10.1021/acs.chemrev.7b00397.

(4) T. Novaes, L. F.; Liu, J.; Shen, Y.; Lu, L.; M. Meinhardt, J.; Lin, S. Electrocatalysis as an Enabling Technology for Organic Synthesis. *Chem. Soc. Rev.* **2021**, *50* (14), 7941–8002. https://doi.org/10.1039/D1CS00223F.

(5) Zahrt, A. F.; Mo, Y.; Nandiwale, K. Y.; Shprints, R.; Heid, E.; Jensen, K. F. Machine-Learning-Guided Discovery of Electrochemical Reactions. *J. Am. Chem. Soc.* **2022**, *144* (49), 22599–22610. https://doi.org/10.1021/jacs.2c08997.

(6) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434–443. https://doi.org/10.1021/acscentsci.7b00064.

(7) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6* (6), 1379–1390. https://doi.org/10.1016/j.chempr.2020.02.017.

(8) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89–96. https://doi.org/10.1038/s41586-021-03213-y.

(9) Jinich, A.; Sanchez-Lengeling, B.; Ren, H.; Harman, R.; Aspuru-Guzik, A. A Mixed Quantum Chemistry/Machine Learning Approach for the Fast and Accurate Prediction of Biochemical Redox Potentials and Its Large-Scale Application to 315 000 Redox Reactions. *ACS Cent. Sci.* **2019**, *5* (7), 1199–1210. https://doi.org/10.1021/acscentsci.9b00297.

(10) Hou, X.; Li, S.; Frey, J.; Hong, X.; Ackermann, L. Machine Learning-Guided Yield Optimization for Palladaelectro-Catalyzed Annulation Reaction. *Chem* **2024**, *0* (0). https://doi.org/10.1016/j.chempr.2024.03.027.
(11) AI4Science, M. R.; Quantum, M. A. The Impact of Large Language Models on Scientific Discovery: A Preliminary Study Using GPT-4. arXiv December 8, 2023. https://doi.org/10.48550/arXiv.2311.07361.
(12) M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting Large Language Models with Chemistry Tools. *Nat. Mach. Intell.* **2024**, *6* (5), 525–535. https://doi.org/10.1038/s42256-024-00832-8.
(13) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous Chemical Research with Large Language Models. *Nature* **2023**, *624* (7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0.
(14) Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **2023**, *145* (32), 18048–18062. https://doi.org/10.1021/jacs.3c05819.
(15) Zheng, Z.; Zhang, O.; Nguyen, H. L.; Rampal, N.; Alawadhi, A. H.; Rong, Z.; Head-Gordon, T.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs. *ACS Cent. Sci.* **2023**, *9* (11), 2161–2170. https://doi.org/10.1021/acscentsci.3c01087.
(16) Rein, J.; Annand, J. R.; Wismer, M. K.; Fu, J.; Siu, J. C.; Klapars, A.; Strotman, N. A.; Kalyani, D.; Lehnherr, D.; Lin, S. Unlocking the Potential of High-Throughput Experimentation for Electrochemistry with a Standardized Microscale Reactor. *ACS Cent. Sci.* **2021**, *7* (8), 1347–1355. https://doi.org/10.1021/acscentsci.1c00328.
(17) Palkowitz, M. D.; Laudadio, G.; Kolb, S.; Choi, J.; Oderinde, M. S.; Ewing, T. E.-H.; Bolduc, P. N.; Chen, T.; Zhang, H.; Cheng, P. T. W.; Zhang, B.; Mandler, M. D.; Blasczak, V. D.; Richter, J. M.; Collins, M. R.; Schioldager, R. L.; Bravo, M.; Dhar, T. G. M.; Vokits, B.; Zhu, Y.; Echeverria, P.-G.; Poss, M. A.; Shaw, S. A.; Clementson, S.; Petersen, N. N.; Mykhailiuk, P. K.; Baran, P. S. Overcoming Limitations in Decarboxylative Arylation via Ag–Ni Electrocatalysis. *J. Am. Chem. Soc.* **2022**, *144* (38), 17709–17720. https://doi.org/10.1021/jacs.2c08006.
(18) Siu, T.; Li, W.; Yudin, A. K. Parallel Electrosynthesis of α-Alkoxycarbamates, α-Alkoxyamides, and α-Alkoxysulfonamides Using the Spatially Addressable Electrolysis Platform (SAEP). *J. Comb. Chem.* **2000**, *2* (5), 545–549. https://doi.org/10.1021/cc000035v.
(19) Kawamata, Y.; Yan, M.; Liu, Z.; Bao, D.-H.; Chen, J.; Starr, J. T.; Baran, P. S. Scalable, Electrochemical Oxidation of Unactivated C–H Bonds. *J. Am. Chem. Soc.* **2017**, *139* (22), 7448–7451. https://doi.org/10.1021/jacs.7b03539.
(20) Horn, E. J.; Rosen, B. R.; Chen, Y.; Tang, J.; Chen, K.; Eastgate, M. D.; Baran, P. S. Scalable and Sustainable Electrochemical Allylic C–H Oxidation. *Nature* **2016**, *533* (7601), 77–81. https://doi.org/10.1038/nature17431.
(21) Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; Zhang, Y. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv April 13, 2023. https://doi.org/10.48550/arXiv.2303.12712.
(22) Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; Kalyan, A. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2507–2521.
(23) Zheng, Z.; Rong, Z.; Rampal, N.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. A GPT-4 Reticular Chemist for Guiding MOF Discovery. *Angew. Chem. Int. Ed.* **2023**, *62* (46), e202311983. https://doi.org/10.1002/anie.202311983.
(24) Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. LLaMA: Open and Efficient Foundation Language Models. arXiv February 27, 2023. https://doi.org/10.48550/arXiv.2302.13971.
(25) Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. Gpt-4 Technical Report. *ArXiv Prepr. ArXiv230308774* **2023**.
(26) Anthropic, A. I. The Claude 3 Model Family: Opus, Sonnet, Haiku. *Claude-3 Model Card* **2024**, *1*.
(27) Xu, F. F.; Alon, U.; Neubig, G.; Hellendoorn, V. J. A Systematic Evaluation of Large Language Models of Code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*; 2022; pp 1–10.
(28) Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. ChemCrow: Augmenting Large-Language Models with Chemistry Tools. arXiv October 2, 2023. https://doi.org/10.48550/arXiv.2304.05376.
(29) Koscher, B. A.; Canty, R. B.; McDonald, M. A.; Greenman, K. P.; McGill, C. J.; Bilodeau, C. L.; Jin, W.; Wu, H.; Vermeire, F. H.; Jin, B.; Hart, T.; Kulesza, T.; Li, S.-C.; Jaakkola, T. S.; Barzilay, R.; Gómez-Bombarelli, R.; Green, W. H.; Jensen, K. F. Autonomous, Multiproperty-Driven Molecular Discovery: From Predictions to Measurements and Back. *Science* **2023**, *382* (6677), eadi1407. https://doi.org/10.1126/science.adi1407.
(30) Ruan, Y.; Lu, C.; Xu, N.; Zhang, J.; Xuan, J.; Pan, J.; Fang, Q.; Gao, H.; Shen, X.; Ye, N.; Zhang, Q.; Mo, Y. Accelerated End-to-End Chemical Synthesis Development with Large Language Models. ChemRxiv May 8, 2024. https://doi.org/10.26434/chemrxiv-2024-6wmg4.
(31) Mahjour, B.; Hoffstadt, J.; Cernak, T. Designing Chemical Reaction Arrays Using Phactor and ChatGPT. *Org. Process Res. Dev.* **2023**, *27* (8), 1510–1516. https://doi.org/10.1021/acs.oprd.3c00186.
(32) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **2022**, *144* (43), 19999–20007. https://doi.org/10.1021/jacs.2c08592.

Table of Contents (TOC) Graphic