Article

Collective intelligence for AI-assisted chemical synthesis

https://doi.org/10.1038/s41586-026-10131-4

Received: 18 July 2025

Accepted: 12 January 2026

Published online: 19 January 2026

Check for updates

Haote Li[1,3], Sumon Sarkar[1,3], Wenxin Lu[1], Patrick O. Loftus[1], Tianyin Qiu[1], Yu Shee[1], Abbigayle E. Cuomo[1], John-Paul Webster[1], H. Ray Kelly[2], Vidhyadhar Manee[2], Sanil Sreekumar[2], Frederic G. Buono[2], Robert H. Crabtree[1], Timothy R. Newhouse[1]✉ & Victor S. Batista[1]✉

The exponential growth of scientific literature presents an increasingly acute challenge across disciplines. Hundreds of thousands of new chemical reactions are reported annually, yet translating them into actionable experiments becomes an obstacle[1,2]. Recent applications of large language models (LLMs) have shown promise[3–6], but systems that reliably work for diverse transformations across de novo compounds have remained elusive. Here we introduce MOSAIC (Multiple Optimized Specialists for AI-assisted Chemical Prediction), a computational framework that enables chemists to make use of the collective knowledge of millions of reaction protocols. MOSAIC is built on the Llama-3.1-8B-Instruct architecture[7], training 2,498 specialized chemical experts in Voronoi-clustered spaces. This approach delivers reproducible and executable experimental protocols with confidence metrics for complex syntheses. With an overall 71% success rate, experimental validation demonstrates the realizations of more than 35 new compounds, spanning pharmaceuticals, materials, agrochemicals and cosmetics. Notably, MOSAIC also enables the discovery of new reaction methodologies that are absent from the expert's training, a cornerstone for advancing chemical synthesis. This scalable model of partitioning vast domains into searchable expert regions enables a generalizable strategy for AI-assisted discovery wherever accelerating information growth outpaces efficient knowledge access and application.

The rapid advancement of science demands efficient methods to navigate an ever-expanding knowledge base. Hundreds of thousands of new reactions are documented annually, joining millions of known transformations across numerous repositories[1,2]. This exponential growth presents a fundamental challenge in which manually accessing information becomes inefficient and expertise-dependent (Supplementary Information Section 19). As chemistry drives innovation in various fields, the ability to translate this knowledge into actionable protocols becomes critical.

The nature of this challenge points to an intriguing solution. As chemistry progresses through iterative experimentation guided by the literature, the field is exceptionally well suited for LLMs integrations. These models, trained on extensive scientific texts, captures intricate relationships underlying chemical concepts, as exemplified by GPT-4 (refs. 8,9). Existing approaches in chemistry have achieved milestone success by developing bespoke models tailored to specific tasks, such as predicting reaction conditions, estimating overall yields or inferring predefined reaction action sequences[10–14]. Recently, there has been notable progress in making use of LLMs as intelligent assistants in chemical research[5]. Systems built on the language-processing capabilities of a generative pre-trained transformer (GPT) have demonstrated the potential to coordinate laboratory automation and synthesis planning[3,4,6].

However, present approaches relying on proprietary models face issues with reproducibility owing to model updates, non-deterministic outputs[15], lack confidence metrics for assessing reliability and struggle with complex chemical inputs in SMILES[4,16] (Supplementary Information Sections 13–17). Although existing bespoke models can suggest reaction conditions or predict yields[17], they fall short in providing the complete details, such as stoichiometry, temperature profiles, and workup steps that determine experimental outcomes. These procedural nuances require extensive laboratory experience, yet even decades of expertise cannot encompass the growing breadth of experimental methodologies. This limitation also affects robotic synthesis platforms, in which missing protocols necessitate frequent human intervention for critical parameters[18,19].

Recent developments in open-access models such as Llama 3.1 (ref. 7) and fine-tuning techniques such as Low-Rank Adaptation (LoRA)[20] have enabled domain specialization for chemistry[21]. We introduce the Multiple Optimized Specialists for AI-assisted Chemical Prediction (MOSAIC) model, a framework that transforms Llama-3.1-8B-Instruct into 2,498 specialized chemistry experts (Fig. 1a–d) using the Pistachio database (see 'Data availability'). This decentralized search-driven approach substantially reduces hardware barriers, enabling subset training on modestly equipped set-ups (4-GPU), rather than large infrastructures requiring tens or hundreds of graphics processing units (GPUs) at

[1]Department of Chemistry, Yale University, New Haven, CT, USA. [2]Chemical Development, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, USA. [3]These authors contributed equally: Haote Li, Sumon Sarkar. ✉e-mail: timothy.newhouse@yale.edu; victor.batista@yale.edu

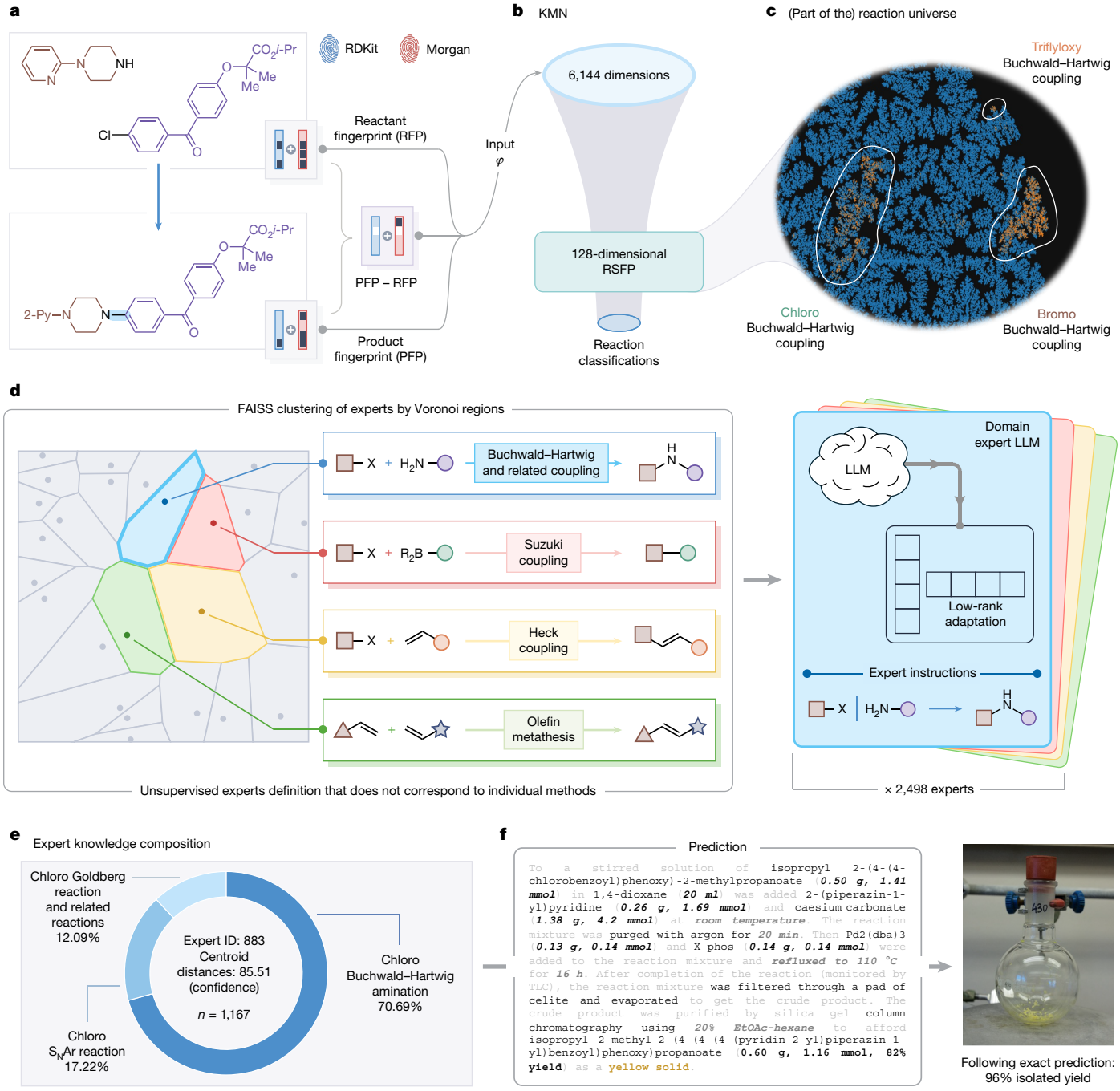


Fig. 1 | MOSAIC framework. a, Buchwald–Hartwig amination reaction fingerprint generation. The reaction components are encoded using concatenated RDKit (blue outline) and Morgan (red outline) fingerprints. A difference fingerprint is computed by subtracting the reactant from product fingerprints, for which black represents +1, white −1 and blank elements 0. **b**, Schematic illustration of the KMN. The input reaction is used by the KMN to generate features and classifications. The feature before the output layer is taken as the RSFP that captures reaction characteristics. **c**, Treemap visualization[45] of the encoded reaction space, highlighting Buchwald–Hartwig reactions (orange) against other reaction classes (blue). The KMN metric effectively distinguishes between chloro, bromo and triflyloxy Buchwald–Hartwig couplings while maintaining intraclass clustering. **d**, Conceptual illustration of specializations in respective Voronoi cells. The reaction universe is clustered into regions by FAISS[46] (each cell has one or several reactions of high similarity) and then fine-tuning Llama-3.1-8B-Instruct with LoRA adaptor on Voronoi domain knowledge. With the domain of related reactions, the LLMs are optimized to produce specialized natural language responses resembling the knowledge seen during training (example shown for Buchwald–Hartwig in consistent blue backgrounds). **e**, Distribution of reaction classes used to train expert 883. This is the top expert for the reaction depicted in **a**, with the Voronoi centroid distance to the query reaction being 85.51. The distribution shows predominant chloro Buchwald–Hartwig amination expertise while maintaining a coverage of related C–N coupling reactions. *n* represents the total number of reactions used to train expert 883. **f**, MOSAIC prediction. A human-reproducible procedure is predicted for this transformation. Details include chemical nomenclature, reagents, solvents, quantities of chemicals, orders of addition, temperature, residence time, workup set-up, product state, overall yield and possible characterization values. 2-Py, 2-pyridyl; *i*-Pr, isopropyl.

once. Moreover, it mitigates hallucinations through specialized expertise, provides quantifiable uncertainty estimation (Fig. 1e and Extended Data Figs. 1 and 2) and enables dynamic scaling for new experts rather than retraining the entire system (Methods).

Using a collection of language models, we achieve fully elaborated, human-readable procedures for chemical synthesis using arbitrary reactions (Fig. 1f). By systematically processing and integrating collective intelligence, this framework offers a valuable tool potentially applicable to the wealth of expanding scientific domains in which expert knowledge must be efficiently accessed and applied.

Quantitative assessment

The development of language models capable of generating comprehensive chemical procedures represents an emerging frontier in synthesis planning. Although previous work has largely focused on specialized models for singular prediction tasks, the ability to interpret and generate end-to-end synthesis procedures, from reagent selection to yield prediction, remains underexplored. Here we introduce quantitative assessments to evaluate how fine-tuned language models handle such tasks. Further, we benchmark MOSAIC with general-purpose LLMs, offering perspectives on chemical understanding in AI systems.

Yield prediction analysis

During predictions, MOSAIC processes the entire experimental procedure, including reagents, solvents and process descriptions, enabling it to anticipate likely experimental outcomes by integrating several dimensions of the synthetic considerations (Fig. 2a).

We implemented a binning strategy that groups yields into ten intervals of ten percentage points from 0 to 100 (specifications in Supplementary Information Section 4). This approach accommodates the token-based nature of the predictions while mitigating experimental variability owing to factors such as individual skill levels and database-reported product impurity.

Our analysis reveals that MOSAIC achieves modest correlations between prediction bin centres and true yield medians ($R^2 = 0.811$; Fig. 2b). Although the tokenization-based approach shows systematic patterns, it incurs substantial prediction errors that limit its quantitative reliability (see the 'Model limitations' section). Also, to ensure the robustness of these findings, we conducted further analysis by limiting reactions to a maximum of 20 instances per class, which yielded comparable performance ($R^2 = 0.809$; Fig. 2c), confirming that the system as a whole captures yield patterns across diverse reaction types rather than memorizing frequent reaction classes.

Reagent and solvent prediction accuracy

To evaluate the accuracy of MOSAIC in predicting reagents and solvents, we used a quantitative metric (D) measuring the overlap between predicted and true molecular sets (Methods and Supplementary Information Section 5). Summarized results are presented in Table 1 and Fig. 2d.

In the simplest case of single predictions (one-shot), the model achieves exact matches for reagents and solvents in 22.4% and 29.8% of cases, respectively, whereas partial matches increase to 45.4% and 51.7%, respectively. When considering predictions using several experts, for reagents, the exact match accuracy nearly doubles to 43.0%, whereas solvent prediction accuracy increases to 32.8%. Statistics for the total number of predictions from the top three experts are shown in Fig. 2e. Furthermore, the partial match success rate in several-expert predictions reaches 76.0% for reagents and 55.2% for solvents. The combined success rate for predicting at least some correct components (reagents or solvents) reaches 94.8%, indicating that MOSAIC almost always identifies relevant reaction components, even if not providing the exact conditions. These results show that consulting several experts improves prediction accuracy.

In cases in which no partial match was achieved even with three experts, our analysis revealed that MOSAIC frequently predicted chemically viable alternatives rather than making erroneous predictions. For example, in nitro-to-amino transformations, the model often predicted iron as a reagent instead of the tin chloride present in the true set. This differentiation reflects the nuanced expertise of the model. In fact, among the top ten experts for such reactions, all focused on nitro-to-amino chemistry, one expert (ranked sixth) specialized in tin chloride transformations (further examples in Supplementary Information Section 5).

Two representative cases illustrating 'half-match' ($D = 0.5$) and 'no-match' ($D = 1$) are analysed in Fig. 2f,g. In the no-match cases, the predictions could represent plausible alternatives to achieve the desired transformations. The D metric represents a lower bound on prediction reliability. When evaluated against literature precedents, it quantifies the guaranteed overlap between predicted and successful reagents and solvents, without penalizing chemically viable alternatives that differ from the reference protocol. This metric is only applicable when literature ground truth exists. Therefore, it cannot assess predictions for new reactions in which protocols are not available. Furthermore, experimental success depends on numerous factors beyond reagent and solvent selection, including procedural details, substrate-specific reactivity and operational parameters. Therefore, direct experimental validations are necessary to assess use in practical synthesis.

Comparison with general-purpose LLMs

To evaluate the capabilities of present language models in addressing chemistry-specific tasks while quantifying the advantage from domain specialization, we compare across diverse and important reaction types. The assessment includes 12 reactions through Suzuki coupling, olefin metathesis, Buchwald–Hartwig amination, Heck reactions, Sonogashira coupling and esterification applied to new substrates with varying complexities. These reactions are tested with: GPT-4o mini, Claude 3.5 Haiku, Claude 3.5 Sonnet and ChatGPT o1 Pro.

The evaluation framework considers scoring criteria related to the chemical understanding and experimental feasibility (Methods). To ensure reliability and account for response variations, we repeat predictions from each model three times using identical prompts that contain the template and example (Fig. 2h). The results, summarized in Fig. 2i and also documented in Supplementary Information Sections 13–18, corroborates the framework's consistent advantage at providing clear instructions for chemical synthesis. Operating with only 8 billion parameters as compared with the likely orders-of-magnitude larger models such as ChatGPT o1 Pro and Claude 3.5 Sonnet, the superior performance of MOSAIC suggests that targeted fine-tuning and chemistry-specific optimization can overcome raw parameter count advantages in specialized domains.

Besides chemistry knowledge, this evaluation revealed that instruction-following capabilities were as important. Models demonstrated markedly different abilities to respond consistently to identical prompts. For example, Claude 3.5 Haiku could provide detailed responses in one trial while refusing to answer in another, claiming insufficient information. This inconsistency presents a challenge for users seeking reliable assistance. Premium models such as ChatGPT o1 Pro and Claude 3.5 Sonnet demonstrated better instruction comprehension and achieved higher scores than GPT-4o mini and Claude 3.5 Haiku, which exhibited erratic behaviour, including superficial template copying and inconsistent response patterns. This variation in instruction-following reliability represents another barrier to using general-purpose LLMs for practical synthesis applications, as chemists

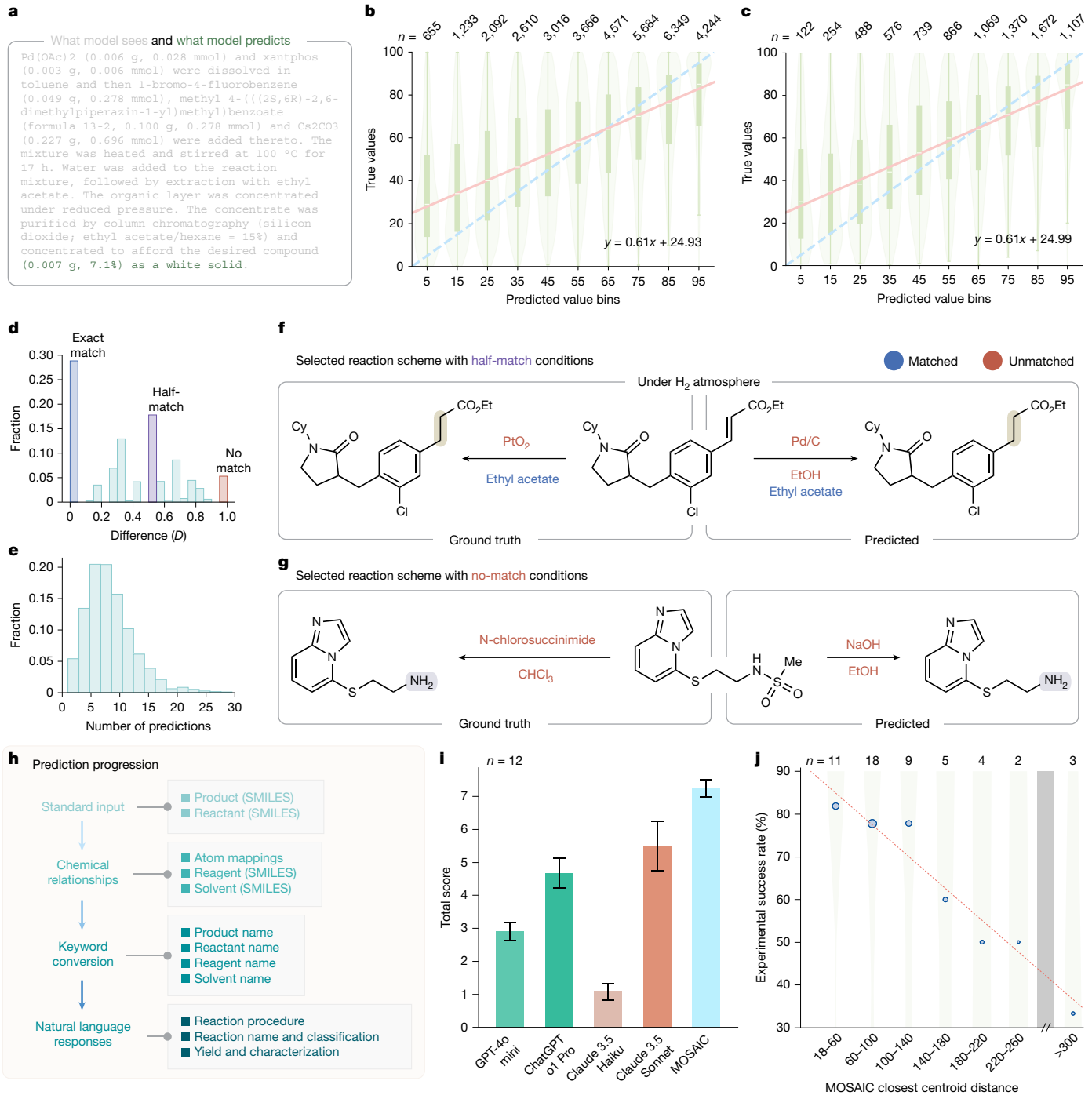


Fig. 2 | Prompt design and quantitative metrics. a, Partial paragraph information processing. The model is provided a partial paragraph. The yield predictions are averaged across beam-searched results. **b**,**c**, Comparative yield prediction analysis showing results with reaction classes capped unconstrained (**b**) and at 20 examples (**c**). *n* represents the sample size. Each plot combines violin and box plots showing true yield distributions within 10% prediction bins. Violin plots show estimates of the full distribution, whereas box plots indicate the 25th and 75th percentiles of the true yields. White lines in each box are medians. Linear fits (solid lines) are plotted against perfect correlation (dashed lines). **d**, Distribution analysis of prediction accuracy. A zero difference (*D*) indicates an exact match. **e**, Frequency distribution of predictions generated by three expert models. **d** and **e** represent 53,227 samples. **f**, Example for half-match. The model suggests palladium on carbon as a catalyst and ethanol as a solvent, typical for alkene hydrogenation reactions. **g**, Example for no match. The model predicts sodium hydroxide in ethanol, a common strategy used for hydrolysis. In both examples, the model suggests reasonable alternatives that could lead to viable synthetic outcomes. **h**, Prompt design. The four-section prompt introduces reactants, completes reagent mappings in SMILES, translates to chemical names and then predicts procedures with classification and yield. **i**, Comparing general-purpose models with MOSAIC on 12 transformations. The result shows superior performance from MOSAIC across the tested reactions. Scores below 5.0 often indicate less useful responses for synthetic practices. Error bars indicate the standard error of the mean. **j**, MOSAIC confidence and experimental success rate plot. Data points represent mean experimental success rate within the binned distance ranges. Sample sizes are indicated by *n*. Green shading represents data distribution. Circle sizes are proportional to *n*. Cy, cyclohexyl; Et, ethyl.

Table 1 | Reagents and solvents prediction results (in % matches)

Prediction type	Match	Reagent	Solvent	Both
One-shot	Exact	22.4	29.8	12.9
	Partial	45.4	51.7	73.0
Several shots	Exact	43.0	32.8	28.9
	Partial	76.0	55.2	94.8

'Both': solvents and reagents as one set.

require consistent and efficient responses without having to troubleshoot model behaviour.

New compounds across broad reactions

To evaluate the practicality, generality and reliability of the proposed framework, we conducted extensive experimental validations by executing exact, highest-ranked predictions on reactions foundational to modern chemical synthesis. These studies examined two notions of novelty: molecular and transformation novelty, which concerns the realizations of previously unreported compounds and reactions. All experiments are detailed in Supplementary Information Section 19 for methodological transparency.

The focus was placed on broadly applicable catalytic reactions central to pharmaceuticals and materials development. The Buchwald–Hartwig amination forms carbon–nitrogen bonds that are ubiquitous among drug molecules. Conditions for these challenging reactions were accurately predicted (Fig. 3 **1a**–**1c**), with notable chemical insight demonstrated through suggestions of palladium-catalysed Buchwald–Hartwig, copper-catalysed Goldberg and S_NAr reactions as viable alternatives across different substrates (see Supplementary Information Section 19, case studies on *N*-arylation reactions). This versatility proved invaluable in synthesizing derivatives of clinically important compounds, including the antidepressant nortriptyline and the cholesterol-lowering medication fenofibrate.

Efficient assembly of drug-like scaffolds was enabled (Fig. 3 **2a**–**2c**), containing sensitive functional groups that typically pose synthetic challenges under Suzuki coupling conditions[22]. The framework was further applied to guide Heck coupling reactions (Fig. 3 **3a**–**3d**), in which previously reported transformations had proved unsuccessful[23] (Fig. 3 **3d**), demonstrating potential to unlock synthetic bottlenecks.

Through application to olefin metathesis reactions (Fig. 3 **4a** and **4b**), precise manipulation of carbon–carbon double bonds in pressure-sensitive adhesive 4-acryloyloxy benzophenone using cross metathesis and in functional material monomers using ring-opening metathesis were achieved, supporting applications from small-molecule synthesis to polymer science[24,25].

Particular strength was demonstrated in alkyne transformations crucial for applications ranging from natural products to functional materials[26]. Navigation of dual-catalytic systems in Sonogashira couplings was achieved (Fig. 3 **5a**), essential for advanced materials and optoelectronic device development[27]. For diaryl ethers, among the most prevalent scaffolds in medicinal chemistry and agrochemicals[28], several viable pathways for estrone derivatives were provided (Fig. 3 **6a** and **6b**), adapting established methods for complex bioactive molecule modification.

Beyond catalytic reactions, transformations in which controlling selectivity and reactivity remains challenging were investigated. Controlled oxidation of pentaerythritol derivatives to corresponding aldehydes was accomplished (Fig. 3 **7a**) while avoiding toxic metals such as chromium[29]. Implication was shown in site-selective and stereoselective reactions of chiral pool materials, including previously unsuccessful[30] conjugate addition to the monoterpene carvone (Fig. 3 **7b**) and stereoselective olefination of *L*-perillaldehyde (Fig. 3 **7c**) using Horner–Wadsworth–Emmons conditions[31].

The capability of the system extended to complex substrate modifications, with prediction of prenylation of the sesquiterpene natural product sclareolide[32] (Fig. 3 **7e**) and site-selective manipulation of multifunctional molecules through silyl monoprotection of accessible phenolic sites of hesperetin (Fig. 3 **7d**), a naturally occurring antioxidant and anti-inflammatory agent. This outcome underscores the ability of the system to capture established selectivity principles from training.

Translational applications

The practical impact of this computational framework spans diverse chemical industries through validated real-world applications. In pharmaceutical development, both new drug-like molecule design (Fig. 4 **8a**) and strategic therapeutic modifications (Fig. 4 **8b**–**8d**) are enabled, proving indispensable for optimizing safety, efficacy and pharmacokinetic properties[33]. In catalysis research, synthesis of specialized ligands for industrial processes[34] (Fig. 4 **9a**) and new photocatalysts that make use of light energy for sustainable chemistry[35] (Fig. 4 **9b** and **9c**) is facilitated. The versatility of the framework extends across materials science, agricultural chemistry and consumer applications. Synthetic routes to conjugated compounds for electronic devices were accurately predicted[36] (Fig. 4 **10a** and **10b**), creation of pyrabactin variants for crop protection was enabled[37] (Fig. 4 **11a**–**11c**) and synthesis of potential fragrance and anti-ageing compounds that are analogues of hedione and retinyl retinoate was achieved[38] (Fig. 4 **12a** and **12b**).

Most notably, MOSAIC has demonstrated potential that enables the development of new methods. As a case study, we used a cascade annulation of heteroaryl dihalides to form bioisosteric analogues of indoles. Conventional annulations of aryl dihalides are well established but analogous transformations for the synthesis of various azaindoles remain underdeveloped[39,40]. Specifically, a study shows that the synthesis of 5-azaindole derivative does not proceed under the existing method[39]. In light of this limitation, MOSAIC guided the development of a new protocol for various azaindole synthesis (Fig. 4 **13a**–**13d**) through an unreported annulation of heteroaryl dihalides with *N*-alkyl allylamines. Notably, 5-azaindole derivative **13d** was afforded under the developed method. For the tasked reaction forming **13a**, the closest expert centroid distance is 320, well above the typical confidence threshold (<150). The lack of closely related precedent reactions underscore that the prediction fell far outside the knowledge space, indicating a genuinely new transformation from the perspective of the predicting expert (details in Supplementary Information Section 19 and Extended Data Figs. 1 and 2). By using the collective knowledge, MOSAIC transforms traditional iterative trial-and-error approaches into informed exploration, accelerating access to previously uncharted regions of chemical space.

Of 37 realized compounds, 35 proceeded on the first attempt using top-ranked predictions, with only two requiring lower-ranked procedures. These successful applications represent most of the examined transformations, although not all predictions yielded successful outcomes. As detailed in Extended Data Figs. 3 and 4, certain reactions resulted in trace yields or fell outside the present capabilities of the model, illustrating the range of experimental outcomes across the evaluated chemical space. Correlation of predicted product colours and physical forms with experimental outcomes is summarized in Supplementary Table 5, with discrepancies found attributable to purity and isolation variations.

Finally, we investigated all synthetic attempts encompassing transformations presented in Figs. 3 and 4, as well as unsuccessful endeavours as shown in Extended Data Figs. 3 and 4. This study enabled determination of the relationship between the closest expert centroid

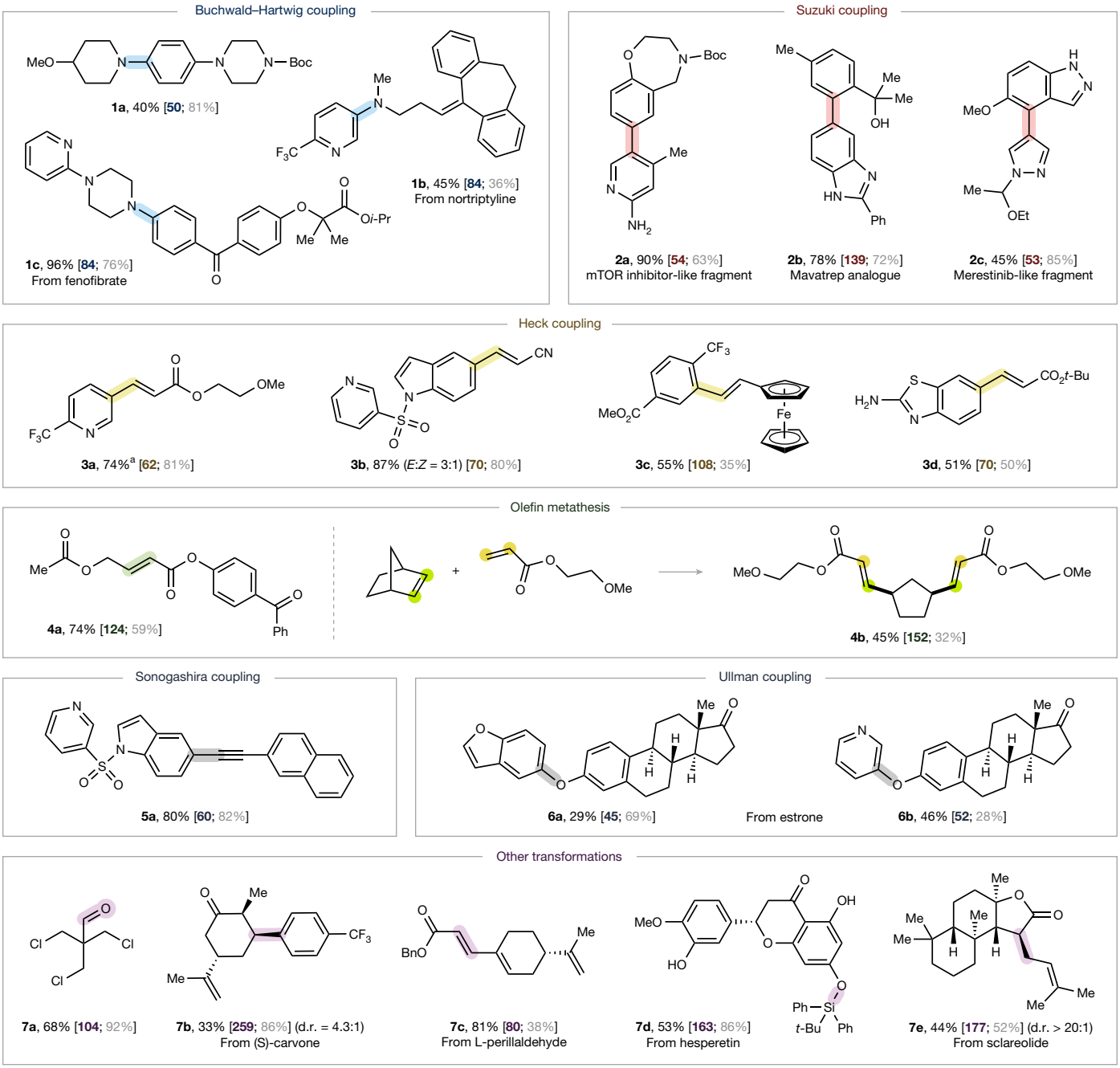


Fig. 3 | De novo compound synthesis guided by computational predictions. New compounds were synthesized using predicted reaction conditions and protocols. Isolated yields are reported for all products. Synthesis was achieved using the highest-ranked predictions unless otherwise indicated. The diverse array of molecular scaffolds demonstrates the capability of the framework to provide reliable synthetic routes for previously unexplored chemical structures across several reaction classes. Bold and grey numbers in square brackets indicate the closest centroid distances to the transformation and MOSAIC-predicted yields, respectively. [a]Second-ranked prediction was successful. Compounds **1a** and **1b**: reported yield from the second-ranked prediction. See Supplementary Information for experimental details. Bn, benzyl; Boc, tert-butyloxycarbonyl; d.r., diastereomeric ratio; Et, ethyl; *i*-Pr, isopropyl; Ph, phenyl; *t*-Bu, tert-butyl.

distances (confidence metric) and experimental outcomes (Fig. 2j). The resulting correlation exposes a predictive framework. Overall, a success rate of 71% was achieved. Transformations positioned in close proximity to expert knowledge domains (distances <100) achieve success rates exceeding 75%, whereas those at greater distances from the corresponding experts yield approximately 50%. This observation establishes a quantitative basis for experimental prioritization, enabling strategic resource allocation between high-confidence targets for more reliable progress and lower-confidence explorations that pave the path to new synthetic methodologies.

Model limitations

Analysis of MOSAIC validation reveals two distinct failure modes that provide strategic guidance for deployment. High-confidence predictions that yield poor experimental results indicate the need for expanded exploration within the prediction space of the model. However, transformations at large distances from the training distribution reflect knowledge boundaries in which databases inadequately capture rapidly evolving methodologies, such as photochemistry (Extended Data Figs. 3 and 4). MOSAIC proves to be most valuable when applied

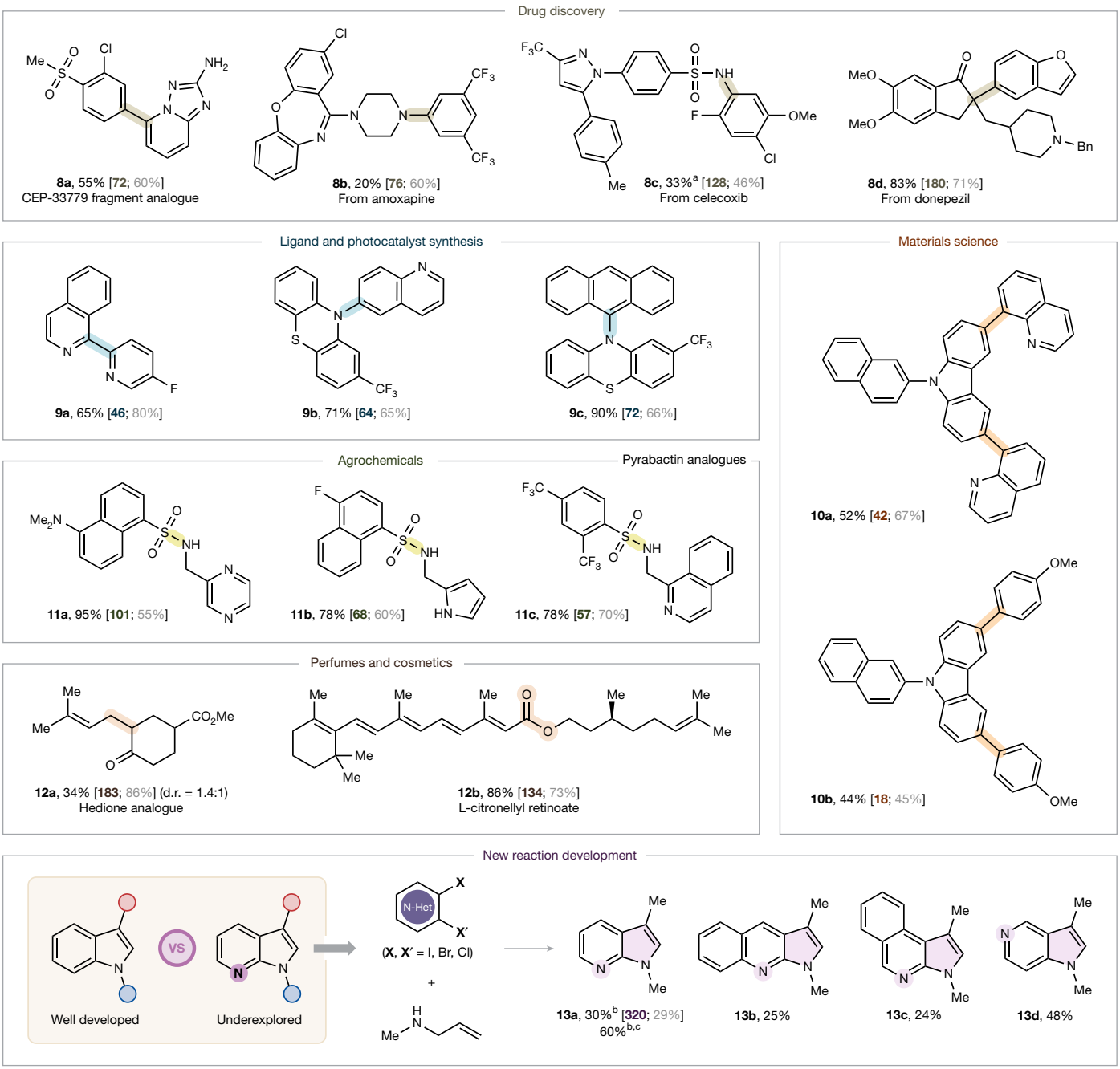


Fig. 4 | Compounds synthesized for translational applications across chemical industries. Isolated yields are reported for all compounds. Products were obtained using the highest-ranked predictions unless otherwise indicated. Synthetic targets span pharmaceutical development, materials science, catalysis research, agricultural chemistry and consumer applications, demonstrating the broad applicability of the computational framework across industrial sectors. Bold and grey numbers in square brackets indicate the closest centroid distances to the transformation and MOSAIC-predicted yields, respectively. [a]Second-ranked prediction was successful. [b]Nuclear magnetic resonance yields. [c]Yield after standard method development practices performed to improve the reaction efficiency. See Supplementary Information for experimental details. Bn, benzyl.

with awareness of these chemical-space limitations, enabling efficient exploration within precedent-rich domains while serving as an honest guide to identify when methodological development may be required.

Moreover, MOSAIC operates within boundaries that reflect the present state of AI in chemistry. The model excels at identifying and adapting known reaction patterns but cannot discover entirely new transformations involving unprecedented reagents, a limitation that connects with the fundamental role of experimental chemistry in advancing new synthetic methodologies. Also, MOSAIC is not expected to achieve optimal yields through one-shot predictions. Successful experimental outcomes in this work establish feasibility towards productive transformations, although iterative optimization may be needed to maximize efficiency for specific substrates.

For specialized applications, the general-purpose architecture of MOSAIC trades precision for breadth. For instance, although it can predict yields for specific Buchwald–Hartwig reactions between 4-methylaniline and aryl halides, it does not match the precision of bespoke models[11] that are optimized for this task using curated datasets[41]. The present implementation inherits constraints from standard LLM tokenization strategies. On the one hand, tokenization inherently

discretizes continuous chemical yield values into finite vocabulary units, causing precision loss. Instead of representing continuous values as real numbers with floating-point precision, LLMs map the infinite space of possible yields (0–100.0%) into a limited set of discrete tokens, fundamentally constraining numerical representation. Moreover, although embedded tokens effectively capture chemical knowledge[42], they are still inefficient with translating between SMILES notation and compound names for large molecules with several heterocyclic rings.

Implementation of chemistry-specific tokenization approaches, such as mixing atom-level encoding[43] or a multimodal representation with explicit molecular graph information[44], could possibly enhance the performance of MOSAIC. Furthermore, although our implementation uses Llama-3.1-8B-Instruct as the base model, the architecture of the framework is model-agnostic and could seamlessly incorporate larger models such as the 70B and 405B models (ref. 7) or the more recent Llama 4 series, which demonstrate superior performance in general language tasks. This flexibility ensures that it can readily incorporate future advances in both language modelling and chemical representation, further narrowing the gap between computational prediction and experimental outcomes.

Discussion

The development of MOSAIC embodies the principle that methods using computational search tend to scale effectively with increasing amounts of data and resources. By partitioning the vast chemical reaction space into searchable Voronoi regions and assigning specialized experts to these regions, MOSAIC can continuously expand its coverage and precision as more data become available. The search mechanism through Facebook AI Similarity Search (FAISS) enables efficient navigation, allowing the system to quickly identify the most relevant expert models for any given query. This architecture allows us to grow the number of experts as new reaction classes emerge. Further, this approach avoids the limitations of strict definitions of reaction types, instead allowing the system to discover and use similarities across transformation patterns directly from the Voronoi cells.

Chemists have already adapted to many changes in how literature is accessed, from physical books to online repositories, and the advent of LLMs offers the next transition. We foresee MOSAIC functioning as a compass in modern chemical synthesis. The integration of LLMs with comprehensive reaction databases creates a powerful in silico platform that enables chemists to systematically obtain reaction procedures and identify viable synthetic routes with speed and precision. The value of MOSAIC lies not in replacing chemical expertise but in rapidly surveying vast chemical space to identify promising experimental directions that would otherwise require extensive literature review and accumulated experience. What once required extensive manual effort and expert intuition for each reaction to determine suitable conditions can now be accomplished within minutes.

Although seasoned chemists may not yet rely on language models for routine transformations, these computational frameworks are becoming increasingly indispensable in contemporary laboratory practice. Chemistry is an empirical science in which new methodologies often emerge through meticulous mechanistic investigation and serendipitous discovery. By integrating empirical and modelling techniques, chemical intuitions and data-driven inspirations are bridged. This approach reduces the time, resources and possible environmental impact associated with reaction development and optimization, while simultaneously expanding the boundaries of the synthetically accessible space.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10131-4.

1. Landhuis, E. Scientific literature: information overload. *Nature* **535**, 457–458 (2016).
2. Llanos, E. J. et al. Exploration of the chemical space and its three historical regimes. *Proc. Natl Acad. Sci. USA* **116**, 12660–12665 (2019).
3. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
4. Bran, A. M. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
5. Ramos, M. C., Collison, C. J. & White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* **16**, 2514–2572 (2025).
6. Zhang, C. et al. SynAsk: unleashing the power of large language models in organic synthesis. *Chem. Sci.* **16**, 43–56 (2025).
7. Dubey, A. et al. The Llama 3 herd of models. Preprint at http://arxiv.org/abs/2407.21783 (2024).
8. White, A. D. The future of chemistry is language. *Nat. Rev. Chem.* **7**, 457–458 (2023).
9. Achiam, J. et al. GPT-4 technical report. Preprint at http://arxiv.org/abs/2303.08774 (2023).
10. Rinehart, N. I. et al. A machine-learning tool to predict substrate-adaptive conditions for Pd-catalyzed C–N couplings. *Science* **381**, 965–972 (2023).
11. Li, S.-W. et al. Reaction performance prediction with an extrapolative and interpretable graph model based on chemical knowledge. *Nat. Commun.* **14**, 3569 (2023).
12. Das, M., Ghosh, A. & Sunoj, R. B. Advances in machine learning with chemical language models in molecular property and reaction outcome predictions. *J. Comput. Chem.* **45**, 1160–1176 (2024).
13. Gao, H. et al. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).
14. Vaucher, A. C. et al. Inferring experimental procedures from text-based representations of chemical reactions. *Nat. Commun.* **12**, 2573 (2021).
15. Chen, L., Zaharia, M. & Zou, J. How is ChatGPT's behavior changing over time? *Harvard Data Science Review* https://hdsr.mitpress.mit.edu/pub/y95zitmz/release/2 (2024).
16. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
17. Tu, Z., Stuyver, T. & Coley, C. W. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chem. Sci.* **14**, 226–244 (2023).
18. Ruan, Y. et al. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nat. Commun.* **15**, 10160 (2024).
19. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
20. Hu, E. J. et al. Lora: low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).
21. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).
22. Zdrazil, B. & Guha, R. The rise and fall of a scaffold: a trend analysis of scaffolds in the medicinal chemistry literature. *J. Med. Chem.* **61**, 4688–4703 (2017).
23. Lu, W. et al. Enhanced ligand discovery through generative AI and latent-space exploration: application to the Mizoroki-Heck reaction. Preprint at https://chemrxiv.org/engage/chemrxiv/article-details/65cfb263e9ebbb4db9859eb7 (2024).
24. Hoveyda, A. H., Malcolmson, S. J., Meek, S. J. & Zhugralin, A. R. Catalytic enantioselective olefin metathesis in natural product synthesis. *Angew. Chem. Int. Ed.* **49**, 34–44 (2010).
25. Sinclair, F., Alkattan, M., Prunet, J. & Shaver, M. P. Olefin cross metathesis and ring-closing metathesis in polymer chemistry. *Polym. Chem.* **8**, 3385–3398 (2017).
26. Gleiter, R. & Werz, D. B. Alkynes between main group elements: from dumbbells via rods to squares and tubes. *Chem. Rev.* **110**, 4447–4488 (2010).
27. Chinchilla, R. & Nájera, C. The Sonogashira reaction: a booming methodology in synthetic organic chemistry. *Chem. Rev.* **107**, 874–922 (2007).
28. Chen, T. et al. Diaryl ether: a privileged scaffold for drug and agrochemical discovery. *J. Agric. Food Chem.* **68**, 9839–9877 (2020).
29. McConnell, J. R., Hitt, J. E., Daugs, E. D. & Rey, T. A. The Swern oxidation: development of a high-temperature semicontinuous process. *Org. Process Res. Dev.* **12**, 940–945 (2008).
30. Bratko, I. et al. Triazolium salts as appropriate catalytic scaffolds for 1,4-additions to α,β-unsaturated carbonyls. *Eur. J. Org. Chem.* **2014**, 2160–2167 (2014).
31. Roman, D., Sauer, M. & Beemelmanns, C. Applications of the Horner–Wadsworth–Emmons olefination in modern natural product synthesis. *Synthesis* **53**, 2713–2739 (2021).
32. Palsuledesai, C. C. & Distefano, M. D. Protein prenylation: enzymes, therapeutics, and biotechnology applications. *ACS Chem. Biol.* **10**, 51–62 (2015).
33. Castellino, N. J., Montgomery, A. P., Danon, J. J. & Kassiou, M. Late-stage functionalization for improving drug-like molecular properties. *Chem. Rev.* **123**, 8127–8153 (2023).
34. Kaes, C., Katz, A. & Hosseini, M. W. Bipyridine: the most widely used ligand. A review of molecules comprising at least two 2,2′-bipyridine units. *Chem. Rev.* **100**, 3553–3590 (2000).
35. Treat, N. J. et al. Metal-free atom transfer radical polymerization. *J. Am. Chem. Soc.* **136**, 16096–16101 (2014).
36. Beaujuge, P. M. & Reynolds, J. R. Color control in π-conjugated organic polymers for use in electrochromic devices. *Chem. Rev.* **110**, 268–320 (2010).
37. Park, S.-Y. et al. Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science* **324**, 1068–1071 (2009).
38. Kim, H. et al. Synthesis and in vitro biological activity of retinyl retinoate, a novel hybrid retinoid derivative. *Bioorg. Med. Chem.* **16**, 6387–6393 (2008).

39. Jensen, T., Pedersen, H., Bang-Andersen, B., Madsen, R. & Jørgensen, M. Palladium-catalyzed aryl amination–Heck cyclization cascade: a one-flask approach to 3-substituted indoles. *Angew. Chem. Int. Ed.* **47**, 888–890 (2008).
40. Fan, R., Wen, H., Chen, Z., Xia, Y. & Fang, W. A general protocol toward synthesis of 3-methylindoles using acenaphthoimidazolyidene-ligated oxazoline palladacycle. *Org. Lett.* **26**, 22–28 (2024).
41. Ahneman, D. T. et al. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
42. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
43. Li, H. et al. Kernel-elastic autoencoder for molecular design. *PNAS Nexus* **3**, 168 (2024).
44. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *Proc. 35th International Conference on Machine Learning* **80**, 2323–2332 (PMLR, 2018).
45. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **12**, 12 (2020).
46. Douze, M. et al. The FAISS library. Preprint at http://arxiv.org/abs/2401.08281 (2025).

Methods

Model framework and design

Training LLMs on extensive datasets poses substantial computational challenges, particularly in coordinating several GPU devices across nodes. Traditional approaches require complex data and model parallelization strategies, alongside intricate synchronization mechanisms[47,48]. For large datasets, conventional non-parallelization-optimized training methods, with limited batch sizes, could extend training times to months for a single investigation.

To overcome this limitation, we build MOSAIC through three progressive components. The first is a distance metric to quantify similarities between chemical reactions. Specifically, we seek a nonlinear kernel function K:

$$K(V_\alpha, V_\beta) \approx \left\| \sum_i [\varphi(V_{(\alpha,i)}) - \varphi(V_{(\beta,i)})]^2 \right\| \quad (1)$$

By design, this function assigns smaller values to similar reactions and larger values to dissimilar ones represented by V. We implemented this idea through a neural network functioning as a nonlinear map φ, in which the Euclidean distance between the pair of transformed reaction descriptions (V_α and V_β) approximates K. This architecture, designated as the Kernel Metric Network (KMN), processes chemical transformations encoded in SMILES notation and classifies them among 2,285 distinct reaction classes (training details in Supplementary Information Section 1). We extract the 128-dimensional representation, termed the reaction-specific fingerprint (RSFP), from the layer preceding the classification heads. RSFP therefore encodes essential information about reaction classifications.

The second component uses the FAISS library for efficient clustering[46]. FAISS implements inverted file indexing[49], which generates quantized Voronoi centroids from vector databases. Although inverted file indexing was originally designed for rapid distance calculations between query vectors and large databases, we repurposed its clustering capability to partition and architect the reaction space. This approach first evaluates distances between query vectors and Voronoi centroids and then for vectors within the selected Voronoi regions. Reactions used during KMN training and validation were encoded as RSFPs, finally constructing a comprehensive chemical transformation space that captures desired metrics. In this space, 2,500 unsupervised Voronoi regions were generated. To develop LLM experts with domain-specific knowledge, we filtered reactions using strict criteria, primarily requiring detailed procedural descriptions (Supplementary Information Section 2). This process resulted in 2,498 non-empty clusters, each being interpreted as a domain of chemical knowledge. The Voronoi-clustering methodology effectively groups related reaction types. For instance, regions containing Buchwald–Hartwig reactions typically encompass related Goldberg and S_NAr reactions (Fig. 1e), demonstrating the systematic ability to recognize chemical similarities.

Finally, the Voronoi cells were used to train domain-specific LLM experts. Rather than initiating independent processes, we first fine-tuned a base model trained on the complete filtered dataset (Supplementary Information Section 2). Subsequently, we continue to individual expert models using data from each set. This approach allows the expert models to maintain diverse knowledge related to chemical nomenclature and substance state characterization while developing specialized expertise in their domains.

When predicting procedures for a new reaction, MOSAIC first encodes the query reaction using KMN to generate the RSFP. This is then used to identify the most relevant regions through FAISS, effectively locating the reaction in the chemical transformation space. For instance, when presented with a Buchwald–Hartwig coupling reaction involving chloro-substituted aromatics (Fig. 1e), the system identified expert 883, whose knowledge composition predominantly consists of chloro Buchwald–Hartwig amination and related C–N-bond-forming transformations. The system activates these domain-specific experts to provide complete synthetic procedures. These detailed guides in natural language are directly executable in laboratory settings. Exactly following the protocol shown in Fig. 1f afforded amination product in 96% isolated yield.

Conditions-matching metric

To evaluate the accuracy of MOSAIC in predicting reagents and solvents, we introduced a quantitative metric (D) to measure the difference between predicted (S_{pred}) and true (S_{true}) sets of molecules:

$$D = 1 - \frac{|S_{pred} \cap S_{true}|}{|S_{true}|} \quad (2)$$

The metric serves two purposes: first, it provides a standardized way to characterize the relationship between predicted and literature conditions; second, it allows systematic categorization of predictions by their degrees of match with reference procedures. This categorization enables identification of cases in which chemically sound alternatives to literature methods exist. A difference of zero ($D = 0$) indicates that the true set is a subset of the predicted set, which is considered as an exact match. Predictions are conducted using both one-shot and several-shots approaches. In the one-shot case, only the first prediction with highest beam score from the top one expert of the query is considered. In the several-shots approach, predictions from up to three experts are aggregated. Furthermore, following an existing approach[50], we record the performance for partial matches, defined as cases in which at least one molecule in the true set appears in the predicted set.

LLM scoring criteria

Models were evaluated using one-shot prompting with a detailed example (Supplementary Information Section 10), an established method for assessing language model performance on domain-specific tasks[51]. Credits are given on the basis of their ability to: correctly map atoms in SMILES notation (1 point), identify appropriate reagents and solvents (1–2 points), detail experimental operation procedures (1 point), specify quantitative parameters such as molar ratios, temperatures and yields (1–2 points), description of workup procedures (1 point) and accurately classify reactions (1 point). Responses that failed to follow instructions or merely replicated the provided examples incurred a penalty of −2 points. This penalty distinguished nonsensical or rote responses from those exhibiting chemical understanding. Each criterion was evaluated independently, rather than binary success/failure assessments. This granular scoring approach allowed us to capture variations in the model capabilities and provide a more nuanced comparison for their chemical-reasoning abilities.

Prompt design

We developed a structured prompt template specifically adapted for chemical contexts. The template logically organizes chemical information to enable reaction prediction and then procedure generation. Our implementation uses the structure described in Fig. 2h. Its design uses the inherent autoregressive nature of transformer–decoder architectures[52] through a carefully arranged sequence to process chemical information. The prediction begins with the processing of the provided product and reactant as primary inputs. From here, the model further generates atom mapping by means of SMILES strings, incorporating both reagents and solvents. The model then derives specific reagents and solvent SMILES from the mappings. Before generating natural language descriptions, the model converts all SMILES notations into standardized chemical names or accepted abbreviations. Using this translated chemical nomenclature, the model synthesizes detailed reaction procedures. Finally, the classification and reaction yields are predicted on the basis of the cumulative information from the previous steps. This

sequential approach enables consistent and chemically meaningful outputs while maintaining the natural flow of information processing.

Experiment prioritization

We define a prioritization scheme that assigns integer ranks starting from 1, with higher values indicating lower priorities for experiments. Given N experts, for which each expert e provides M_e predictions (M_e varies by expert), we establish a ranking function $R(e,p)$, in which $e \in [1, N]$ represents the expert index and $p \in [1, M_e]$ denotes the prediction index for expert e. The ranking is determined by:

$$R(e, p) = N(p - 1) + e$$

This formulation ensures a systematic ordering in which all predictions at priority level p are ranked before proceeding to level $p + 1$, while maintaining a consistent ordering among experts within each priority level. When an expert has exhausted their predictions, they no longer contribute to subsequent priority levels. Applications of the prioritization strategy are provided in Supplementary Information Section 19.

Incremental expert scaling

MOSAIC supports incremental updates through a hierarchical indexing design that avoids retraining the entire framework and requires no rebalancing of existing clusters. When new reaction data become available, the original FAISS index with its 2,498 Voronoi regions and trained experts is preserved, whereas a second-level index is created for the new reactions. The existing KMN generates RSFPs, which are then clustered into further Voronoi regions in the second index. New experts are trained on these clusters following the same progressive fine-tuning approach, starting from the base model trained on the complete filtered dataset. During prediction, the search operates hierarchically: results are retrieved first from the original index, then from the second-level index, concatenated and ranked by centroid distance to select the most relevant experts. This process can be repeated iteratively as more data become available. Example code and annotated notebooks are available ('Code availability').

Safety guidelines

All chemical procedures produced by MOSAIC must only be carried out by individuals with appropriate safety training and in properly equipped laboratory environments. Many chemical reactions involve hazardous materials, potentially dangerous conditions or risks that may not be fully detailed in the procedural descriptions. Safe and successful execution requires thorough knowledge of chemical reactivity and strict adherence to safety protocols. We acknowledge that no AI system, including MOSAIC, can yet guarantee absolute safety across all queries. Users should be aware that predictions vary in reliability and should critically evaluate all suggestions before implementation. The confidence metrics provided by MOSAIC serve as the first layer of assessment, indicating the familiarity of the model with similar reaction types. For reactions with high confidence scores, users can review the evidence trails provided by reaction references to verify alignment with established literature practices. For low-confidence predictions, especially those involving unexplored chemical space, further verification through traditional literature searches is recommended. MOSAIC is designed to enable queries to the vast foundational knowledge in the most tangible ways but not to replace chemical intuition and established reference sources. The system's transparent confidence assessments and reference trails are intended to foster responsible implementation rather than encouraging uncritical acceptance.

Data availability

The Pistachio database (version number 2024Q1) can be accessed at https://www.nextmovesoftware.com/pistachio. Source data are provided with this paper.

Code availability

The code and annotated notebooks are available at https://github.com/haoteli/MOSAIC. An archived code base and the reaction universe visualization are available from Zenodo at https://doi.org/10.5281/zenodo.17904274 (ref. 53).

47. Zhao, Y. et al. PyTorch FSDP: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.* **16**, 3848–3860 (2023).
48. Rajbhandari, S. et al. ZeRO-infinity: breaking the GPU memory wall for extreme scale deep learning. In *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis* 59 (ACM, 2021).
49. Zobel, J. & Moffat, A. Inverted files for text search engines. *ACM Comput. Surv.* **38**, 6-es (2006).
50. Andronov, M. et al. Reagent prediction with a molecular transformer improves reaction data quality. *Chem. Sci.* **14**, 3235–3246 (2023).
51. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
52. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
53. Li, H. MOSAIC: Multiple Optimized Specialists for AI-assisted Chemical Prediction. *Zenodo* https://doi.org/10.5281/zenodo.18002953 (2025).
54. Sarkar, S., Ghosh, S., Kurandina, D., Noffel, Y. & Gevorgyan, V. Enhanced excited-state hydricity of Pd–H allows for unusual head-to-tail hydroalkenylation of alkenes. *J. Am. Chem. Soc.* **145**, 12224–12232 (2023).
55. Bodnar, A. K. & Newhouse, T. R. Accessing *Z*-enynes via cobalt-catalyzed propargylic dehydrogenation. *Angew. Chem. Int. Ed.* **63**, e202402638 (2024).
56. Geunes, E. P., Meinhardt, J. M., Wu, E. J. & Knowles, R. R. Photocatalytic anti-Markovnikov hydroamination of alkenes with primary heteroaryl amines. *J. Am. Chem. Soc.* **145**, 21738–21744 (2023).
57. Ratushnyy, M., Kvasovs, N., Sarkar, S. & Gevorgyan, V. Visible-light-induced palladium-catalyzed generation of aryl radicals from aryl triflates. *Angew. Chem. Int. Ed.* **59**, 10316–10320 (2020).
58. Xie, K. A. et al. A unified method for oxidative and reductive decarboxylative arylation with orange light-driven Ir/Ni metallaphotoredox catalysis. *J. Am. Chem. Soc.* **146**, 25780–25787 (2024).

Acknowledgements We acknowledge support from Boehringer Ingelheim. V.S.B. also acknowledges a generous allocation of high-performance computing time from the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under contract no. DE-AC0205CH11231 using NERSC awards BES-ERCAP0030682, BES-ERCAP0024372 and BESERCAP0027329 and partial support for the development of the MOSAIC software from the National Science Foundation Engines Development Awards: Advancing Quantum Technologies (CT) under award number 2302908. We acknowledge Yale University's Chemical and Biophysical Instrumentation Center (CBIC) and Catalysis and Separations Core (CSC) for providing analytical resources and F. Menges for obtaining the high-resolution mass spectrometry data. H.L. thanks discussions with J. E. Crivelli-Decker and B. J. Shields et al. at SandboxAQ.

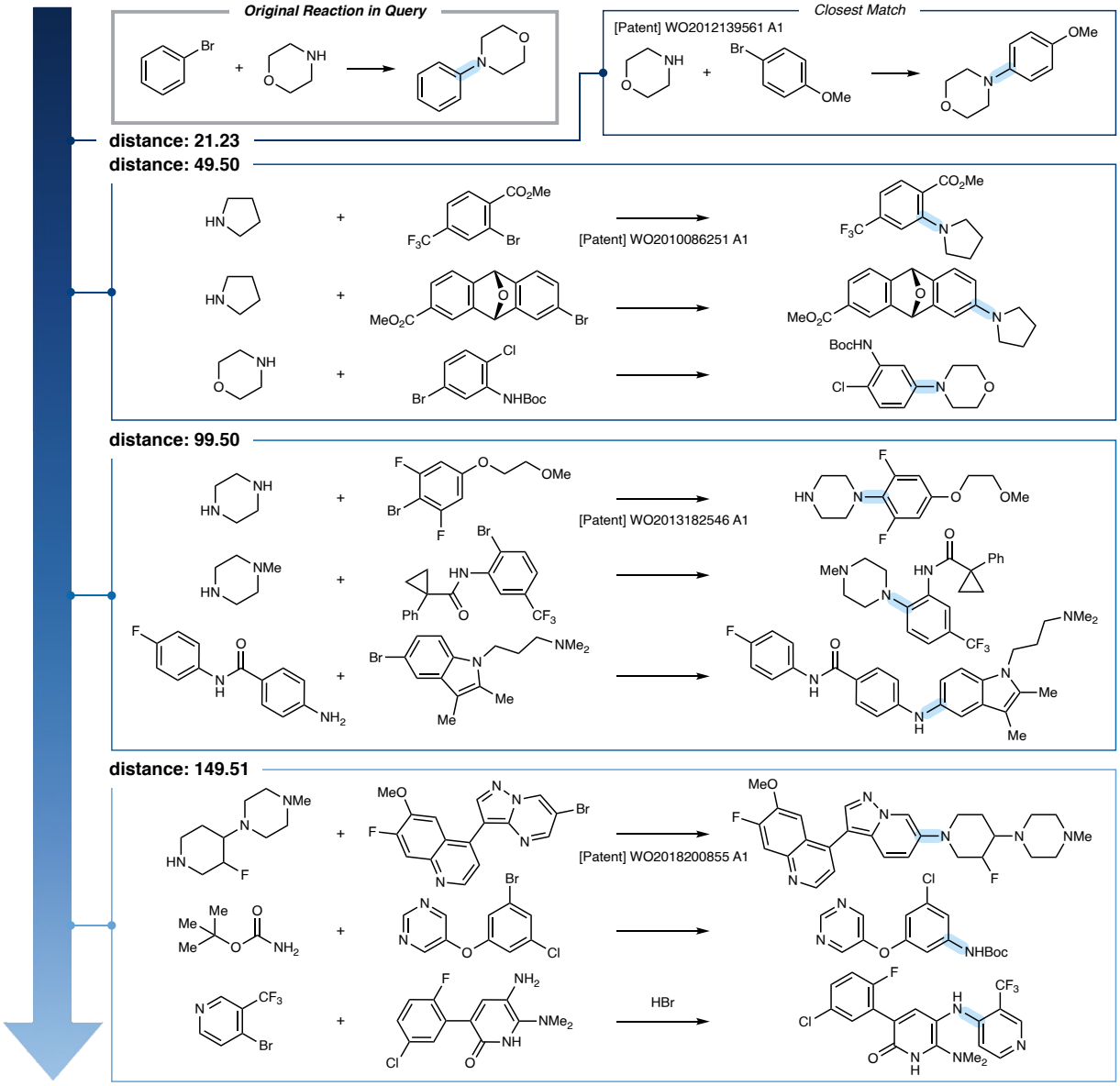
Author contributions H.L. conceived the method, developed the theoretical framework, implemented the computational applications and wrote the initial manuscript. S. Sarkar designed and performed experimental validations, defined and examined the application scope and contributed to model improvements. W.L. contributed key figure illustrations, experimental designs and validation studies. P.O.L. conducted experimental validations that strengthened the conclusions. T.Q. contributed figure illustrations. Y.S. processed the Pistachio dataset. A.E.C., J.-P.W., H.R.K., V.M., S. Sreekumar and F.G.B. advised the method. T.R.N. and V.S.B. acquired funding. T.R.N., V.S.B. and R.H.C. supervised this research and all authors revised the manuscript for publication. H.L., S. Sarkar and W.L. jointly prepared the final manuscript.

Competing interests The authors declare no competing interests.

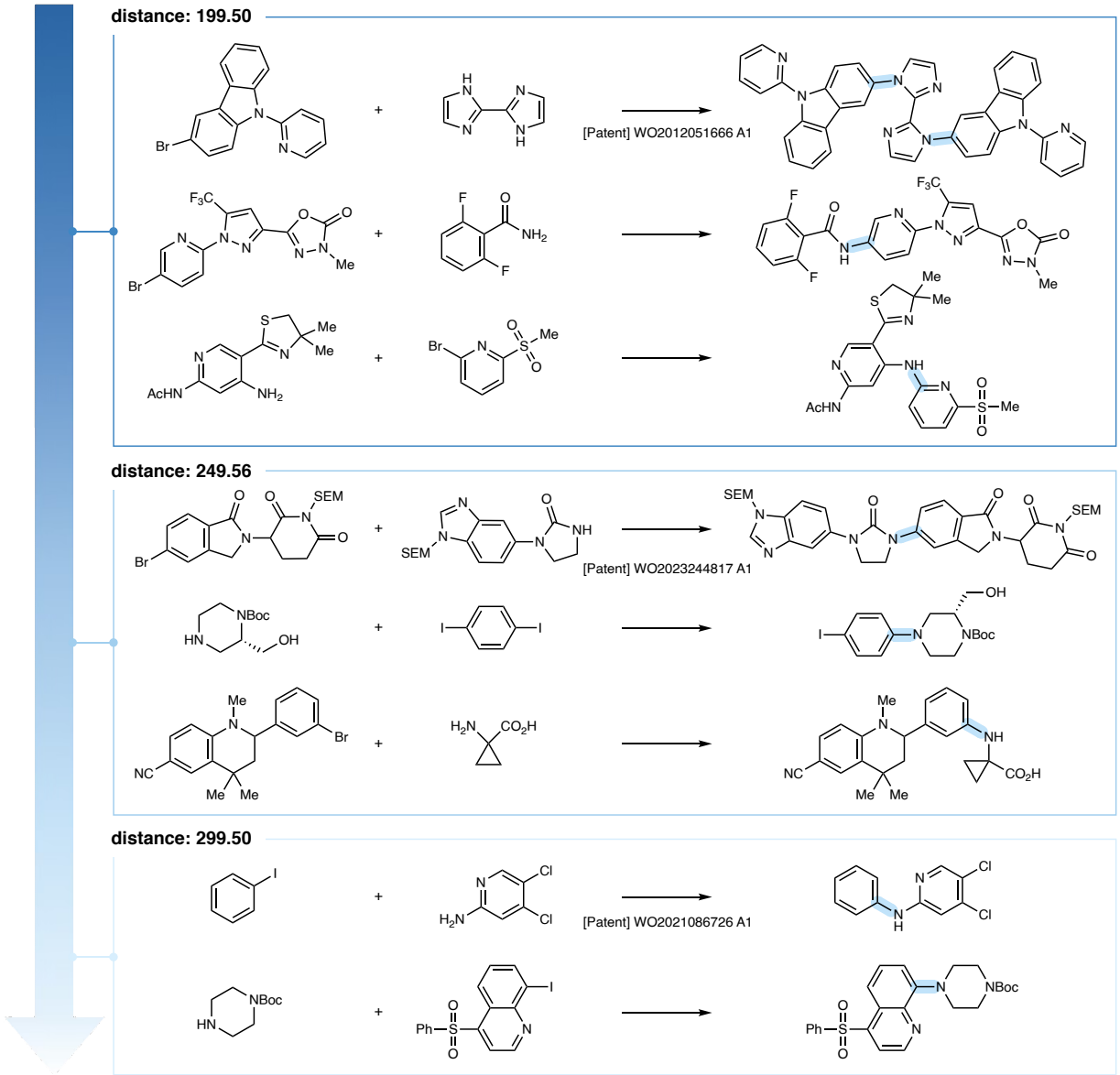
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10131-4.

Correspondence and requests for materials should be addressed to Timothy R. Newhouse or Victor S. Batista.

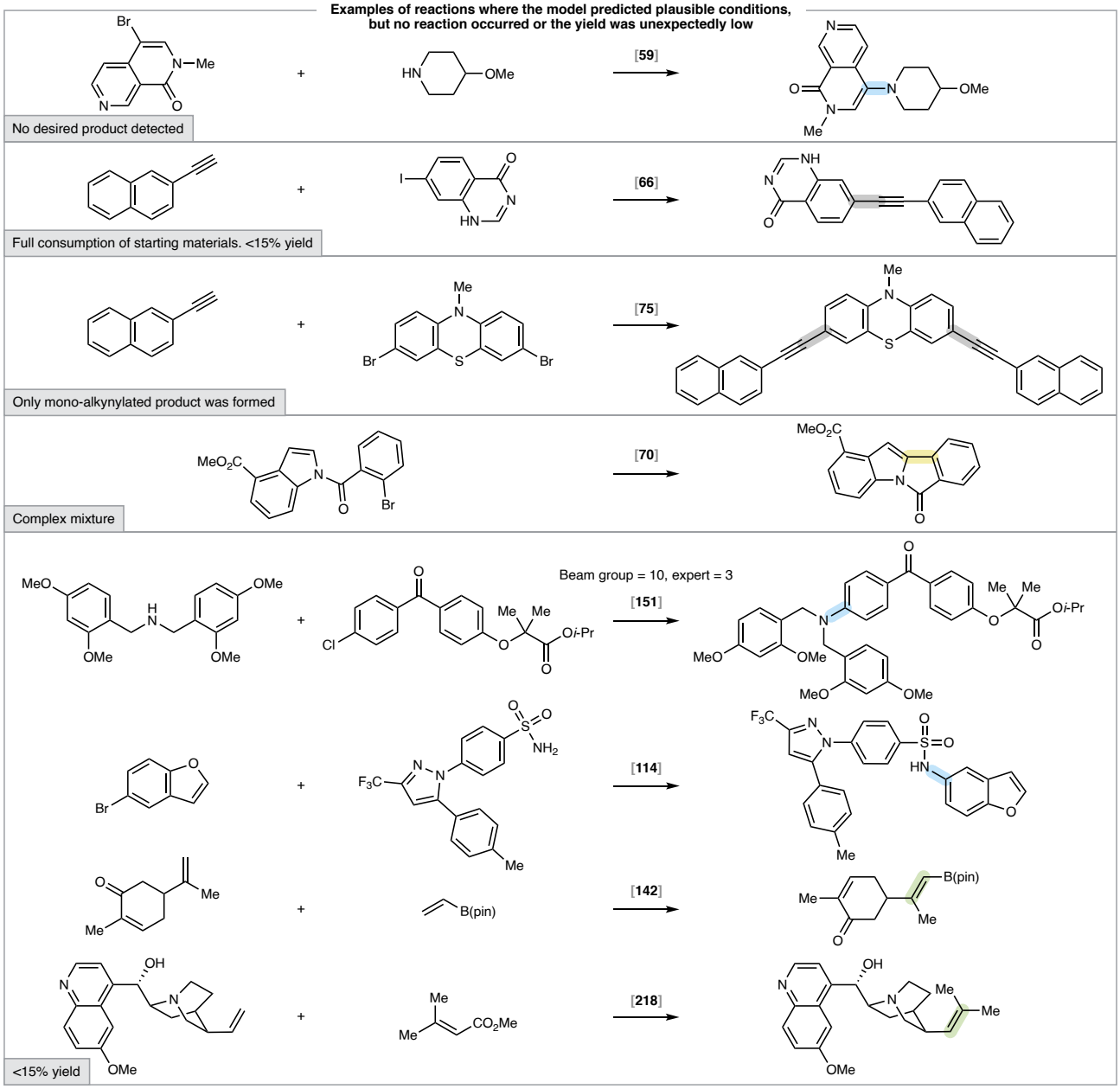
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



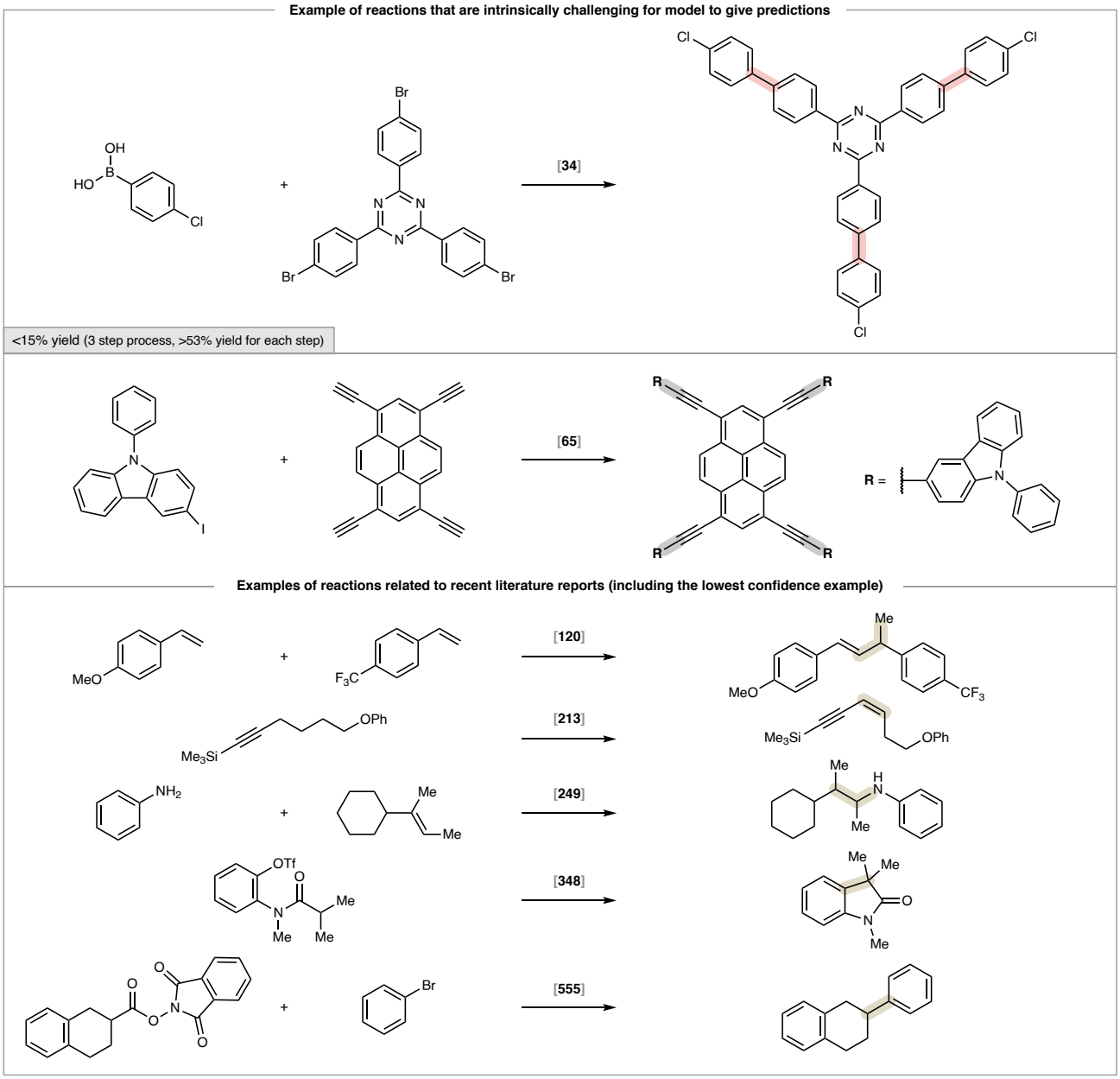
Extended Data Fig. 1 | Higher-confidence examples. Analysis of confidence (distance) patterns reveals distinct thresholds that correlate with prediction reliability. For example, high-confidence predictions (distances <50) demonstrate strong structural and mechanistic resemblance, sharing similar transformation patterns with close similarity in both reactant and product structures. Moderate confidence predictions (distances 100–200) retain core transformation patterns while exhibiting greater variation in substrate and product structures. Boc, tert-butyloxycarbonyl.



Extended Data Fig. 2 | Lower-confidence examples. Lower-confidence predictions (distances >200) fall in the same broad reaction category but often involve different reactive groups or alternative reaction conditions. These predictions provide valuable synthetic insights. When querying the Buchwald–Hartwig coupling of aryl bromides at distance 300, analogous couplings with aryl iodides were identified, suggesting alternative synthetic approaches or serving as mechanistic inspiration for method developments. Boc, tert-butyloxycarbonyl; SEM, 2-(trimethylsilyl)ethoxymethyl.



Extended Data Fig. 3 | Failed examples with higher confidence. There are cases in which MOSAIC showed high confidence (distance <100) but yielded poor experimental results (<15% yield) with top two predicted procedures. Despite chemically sound protocols representing appropriate reaction classes, these transformations require expanded prediction exploration beyond initial top-ranked suggestions to achieve practical synthetic success. *i*-Pr, isopropyl; pin, pinacolato.



Extended Data Fig. 4 | Examples with reactions existing at distance from MOSAIC's training distribution. Representative transformations require specialized methodologies poorly captured in present patent-dominated databases. These cases highlight the need for incorporating more experts from rapidly evolving fields such as photochemistry to expand model capabilities into disconnected chemical space[54–58]. Despite showing high confidence with the transformation, the model seemed to struggle with generation for symmetric multiple functionalization in which no valid predictions were produced from the top two experts. Ph, phenyl; Tf, trifluoromethanesulfonate.