COLLECTIVE INTELLIGENCE OF SPECIALIZED LANGUAGE MODELS GUIDES REALIZATION OF DE NOVO CHEMICAL SYNTHESIS

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

[1]Department of Chemistry, Yale University, New Haven, Connecticut 06511, USA
[2]These authors contributed equally
[3]Chemical Development, Boehringer Ingelheim Pharmaceuticals Inc, Ridgefield, Connecticut 06877, USA
[*]Corresponding authors: timothy.newhouse@yale.edu and victor.batista@yale.edu

ABSTRACT

While hundreds of thousands of new chemical reactions are reported annually, efficient use of this vast collection of synthetic knowledge remains a persistent challenge in modern chemistry. Recent applications of large language models (LLMs) have shown promise, but systems that reliably work for de novo compounds and molecular transformations have remained elusive. Here we introduce MOSAIC (Multiple Optimized Specialists for AI-Driven Chemical Prediction), a computational framework that enables chemists to harness the collective knowledge of millions of reaction protocols. In contrast to existing approaches relying on agentic models, MOSAIC leverages the open-source Llama-3.1-8B-instruct architecture. By training 2,489 specialized chemical experts on Voronoi-clustered reaction spaces, we establish a scalable paradigm that delivers reproducible and human-readable experimental protocols for complex syntheses. Experimental validation demonstrates MOSAIC's ability to predict and execute previously unreported transformations, including challenging reactions via Buchwald-Hartwig amination, Suzuki coupling, and olefin metathesis. We validate this approach through the successful synthesis of over 35 novel compounds spanning pharmaceuticals, materials, agrochemicals, and cosmetics. This framework establishes a new relationship between computational and experimental chemistry, providing a foundation for accelerated chemical discovery across disciplines.

Main

The rapid advance of chemical sciences demands efficient methods to navigate and utilize the vast and ever-expanding repository of synthetic knowledge. Every year, hundreds of thousands of new chemical reactions are documented, adding to millions of known transformations dispersed across numerous repositories. Traditional approaches to reaction planning often involve manual searches of the literature or databases to identify procedures using structurally similar starting materials and products. This process is time-consuming, highly dependent on individual expertise, and introduces significant inefficiencies, creating a bottleneck in chemical discovery and development. As the field of chemistry expands into interdisciplinary areas such as materials science, pharmaceuticals, and sustainable technologies, the challenge of effectively accessing and applying this growing body of knowledge become increasingly acute.

The nature of this challenge points to an intriguing solution: as chemistry essentially progress through iterative synthetic experimentation guided by insights from the chemical literature, the field is exceptionally well-suited for the application of large language models (LLMs). These models, trained on extensive collections of scientific texts, can capture the intricate relationships and contextual meanings underlying chemical concepts, as exemplified by systems such as GPT-4 [1, 2].

Traditional AI approaches in chemistry have achieved milestone success by developing bespoke models tailored to specific tasks, such as predicting reaction conditions [3], estimating overall yields [4, 5], or inferring pre-defined reaction action sequences [6]. Recently, there has been remarkable progress in leveraging LLMs as intelligent assistants in chemical research [7]. Systems such as Coscientist and ChemCrow [8, 9, 10], built on GPT-4's advanced natural language processing capabilities, have demonstrated the potential to coordinate laboratory automation and guide synthesis planning, an exciting step toward AI-assisted chemistry.

However, these early efforts have revealed fundamental challenges that must be addressed for efficient AI-driven chemical synthesis. A central limitation of approaches that rely on proprietary models such as GPT-4 is the dependence on closed-source systems whose architectures or configurations can yet to be modified for chemistry-specific tasks. These models generate different responses to identical prompts across different sessions [11] (Supplementary Section 11–15), compromising experimental reproducibility. The absence of confidence metrics in these systems make non-expert users unable to effectively interpret the reliability of the outputs. While such methods perform adequately with simple, well-studied substances like aspirin or benzoic acid, their utility diminishes with more complex or novel molecules requiring standarized SMILES inputs [12] (Supplementary Section 11–15). Furthermore, while existing combinations of non-LLM models can suggest relevant reaction conditions, they fall short in providing details necessary for experimental implementation. Human experts are often required to make iterative decisions [13] or needed to manually determine critical parameters such as concentration equivalence ratios, stoichiometry, order of addition, and residence times [14], creating a persistent bottleneck in both automated and manual experimental workflow.

Addressing these challenges requires a paradigm shift in how we apply language models to chemistry. Recent developments in open-access models, such as Llama-3.1 [15], and efficient fine-tuning techniques such as the Low-Rank Adaptation (LoRA) [16], have opened new possibilities for domain specialization in chemistry [17]. Building on these advances, we introduce the Multiple Optimized Specialist for AI-driven Chemical Prediction (MOSAIC) model, a framework that transforms the Llama-3.1-8B-instruct model into 2,489 specialized chemistry experts using the Pistachio database [18], harnessing massively scalable computational power.

MOSAIC's architecture enables the optimization of individual models on partitioned subsets of chemical reactions to develop domain-specific expertise across a broad range of chemical transformations. By directly utilizing SMILES strings as inputs, the expert models generate detailed, human-reproducible reaction procedures, including precise reagent and solvent stoichiometry information, along with yield predictions. A distinctive feature of this approach is the inclusion of confidence estimates, achieved by quantifying the transformation distance from the query to the domain of expertise, providing essential reliability metrics. We experimentally validate these capabilities on de novo compounds across pharmaceuticals, materials, and cosmetics and agrochemical applications, demonstrating real-world practicality. The framework is designed to scale continuously with increasing data volume and computational resources. Through the parallel fine-tuning of thousands of models and the generation of complete experimental protocols, our model represents a significant step to create an integrated, all-in-one system for synthesis and chemical discovery.

To the best of our knowledge, this is the first demonstration of using a collection of large language models to achieve fully elaborated, human-readable procedures for de novo chemical compound synthesis employing arbitrary reactions. The framework's systematic approach to searching a reaction space and providing confidence-aware predictions establishes it as a valuable tool for accelerating discovery across the chemical sciences. By combining the wealth of knowledge from the expanding chemical literature with the reasoning capabilities of foundation language models, MOSAIC empowers chemists to focus on idea generation while leveraging machine intelligence to optimize experimental outcomes.

Model Framework and Design Logics

Training large language models on extensive datasets presents significant computational challenges, particularly in coordinating multiple GPU devices across nodes. Traditional approaches require complex data and model parallelization strategies, alongside intricate synchronizations mechanisms [20, 21]. For datasets of this magnitude, conventional non-parallelization-optimized training methods, with limited batch sizes, could extend training times to several months for a single investigation.

To overcome these limitations, we equipped MOSAIC with three distinct architectural components (Fig. 1**a**). The first component implements a distance metric to quantify similarities between chemical reactions. Specifically, we developed a non-linear kernel function $\mathcal{K}$:

$$\mathcal{K}(\vec{V}_1, \vec{V}_2) \approx \left\| \sum_i (\varphi(\vec{V}_{1,i}) - \varphi(\vec{V}_{2,i}))^2 \right\| \qquad (1)$$

By design, this function assigns smaller values to similar reactions and larger values to dissimilar ones represented by $\vec{V}$. We implemented this idea through a neural network functioning as a non-linear map φ, where the Euclidean distance between the pair of transformed reaction descriptions ($\vec{V}_1$ and $\vec{V}_2$) approximates $\mathcal{K}$. 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 Section 1). We extract

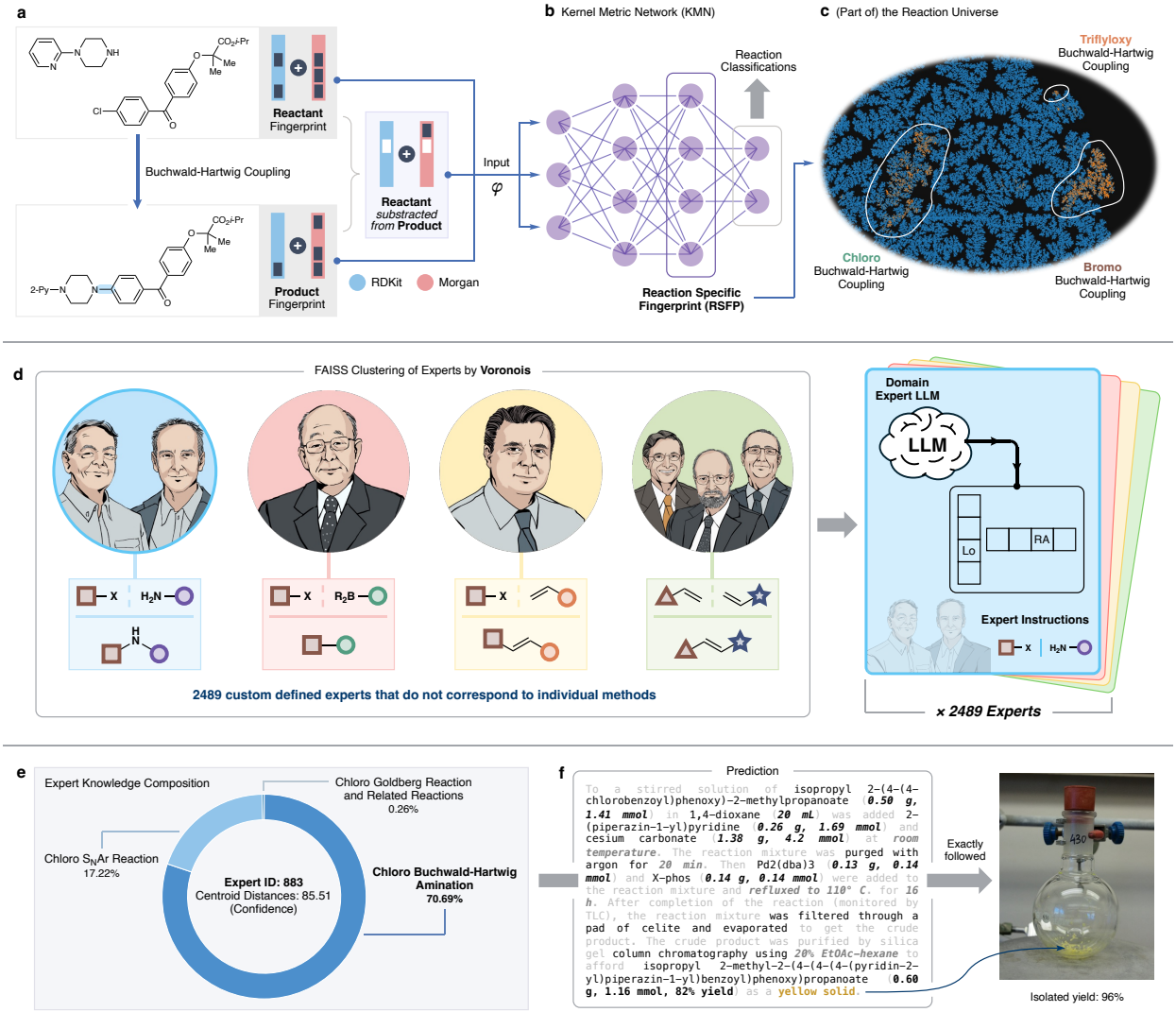

Figure 1: **MOSAIC framework**. **a**, Buchwald-Hartwig amination reaction fingerprint generation. The reaction components are encoded using concatenated RDKit (blue outline) and Morgan (red outline) fingerprints. A difference fingerprint is computed by subtracting the reactant from product fingerprints, where black represents +1, white -1, and outlined elements 0. **b**, schematic illustration of the KMN. The input from the reaction is used by the KMN to classify reaction classes during training. The feature before the output layer is taken as the Reaction Specific Fingerprint that captures relevant reaction characteristics. **c**, Tree Map visualization [19] of RSFP-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 intra-class clustering. **d**, Conceptual illustration of Voronois of experts. The reaction universe is clustered into discrete Voronoi cells by FAISS. Each Voronoi cell have one or several reactions of high similarity. And then training Meta's Llama-3.1-8B-Instruct with LoRA adapter using Voronoi domain knowledge. With the domain of related reactions, the LLM models are optimized to produce specialized natural language responses resembling the knowledge seen during training (example shown for Buchwald-Hartwig in consistent light blue background). **e**, distribution of reaction classes used to train expert 883. This is the top expert for reaction depicted in **a** with the expert's Voronoi centroid distance to the query reaction being 85.51. The distribution shows predominant Chloro Buchwald-Hartwig amination expertise while maintaining a well-rounded coverage of related C–N coupling reactions. **f**, the prediction from model for the novel compound. The model predicts a human-reproducible procedure for this transformation. Details include chemical nomenclature, reagents, solvents, quantitative ratio between each chemical, orders of addition, temperature, residence time, workup setup, product state, overall yield, and possible characterization values. *i*-Pr, isopropyl; 2-Py, 2-pyridyl.

the 128-dimensional feature vector—termed the Reaction-Specific Fingerprint (RSFP) from the fully connected layer preceding the classification heads. RSFP encodes essential information about reaction classifications.

Clustering with FAISS and Voronoi Regions: The second component leverages the Facebook AI Similarity Search (FAISS) library for efficient clustering[22]. FAISS implements Inverted File Indexing (IVF)[23], which generates quantized Voronoi centroids from vector databases. While IVF was originally designed for rapid distance calculations between query vectors and large databases, we repurposed its clustering capability to partition the reaction space into Voronoi regions. This approach first evaluates distances between query vectors and Voronoi centroids, then computes detailed distances only for vectors within the selected Voronoi regions. Reactions processed during KMN training and validation were encoded as RSFPs, constructing a comprehensive chemical transformation space that captures reaction similarities. Within this space, 2,500 self-supervised clustered Voronoi regions were generated. To develop LLM experts with domain-specific knowledge, we filtered reactions using strict criteria, primarily requiring detailed procedural descriptions (Supplementary Section 2). This filtering process resulted in 2,489 non-empty Voronoi regions, each representing a domain of chemical knowledge. The Voronoi-clustering methodology effectively groups related reaction types. For instance, Voronoi regions containing Buchwald-Hartwig reactions typically encompass related Goldberg and S_NAr reactions (Fig. 1**e**), demonstrating its systematic ability to recognize chemical similarities.

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

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

Quantitative Assessment

The development of language models capable of generating comprehensive chemical procedures represents an emerging frontier in synthesis planning. While prior work has largely focused on specialized models tailored to specific prediction tasks, the ability to interpret and generate end-to-end synthesis procedures, from reagent selection to yield prediction, remains a nascent area of research. Here, we introduce quantitative assessments for evaluating how fine-tuned language models handle the complex task of detailed synthesis procedures. We benchmark MOSAIC against leading language models, including the ones that are publicly accessible and subscription-based, offering new perspectives on chemical reasoning in AI systems.

Yield Prediction Analysis

We first examined the model's ability to predict reaction yields based on complete experimental procedures. LLMs represent yields as tokens rather than continuous numerical values, a different perspective on yield prediction compared to traditional approaches. During predictions, MOSAIC processes the entire experimental procedure, including reagents, solvents, and process descriptions, enabling it to anticipate likely experimental outcomes by integrating multiple dimensions of the synthetic considerations (Fig. 2**c**).

Binning Strategy for Token-Based Predictions: To evaluate yield predictions, we implemented a binning strategy that groups yields into 10 intervals of 10 percentage points from 0 to 100 (specifications in Supplementary Section 4). This approach not only accommodates the token-based nature of the predictions while mitigating experimental variability due to factors such as individual skill levels and database-reported product impurity. Within each bin, we compared the medians of the interval against the average of corresponding true values within the bin.

Correlation and Robustness: Our analysis reveals that MOSAIC demonstrates a qualitative understanding of yield trends, evidenced by the correlation between predicted and true yield medians (Fig. 2**e**). The model's ability to integrate information from complete experimental procedures provides a complementary approach to specialized yield prediction models. To ensure the robustness of these findings, we conducted additional analysis by limiting a maximum of 20 instances per reaction class, confirming that the system captures yield patterns across diverse reaction types (Fig. 2**d**).

Reagent and Solvent Prediction Accuracy

To evaluate the accuracy of MOSAIC in predicting reagents and solvents (detailed in Supplementary Section 9), we introduced a quantitative metric (D) to measure the difference between predicted (S_{pred}) and true (S_{true}) sets of

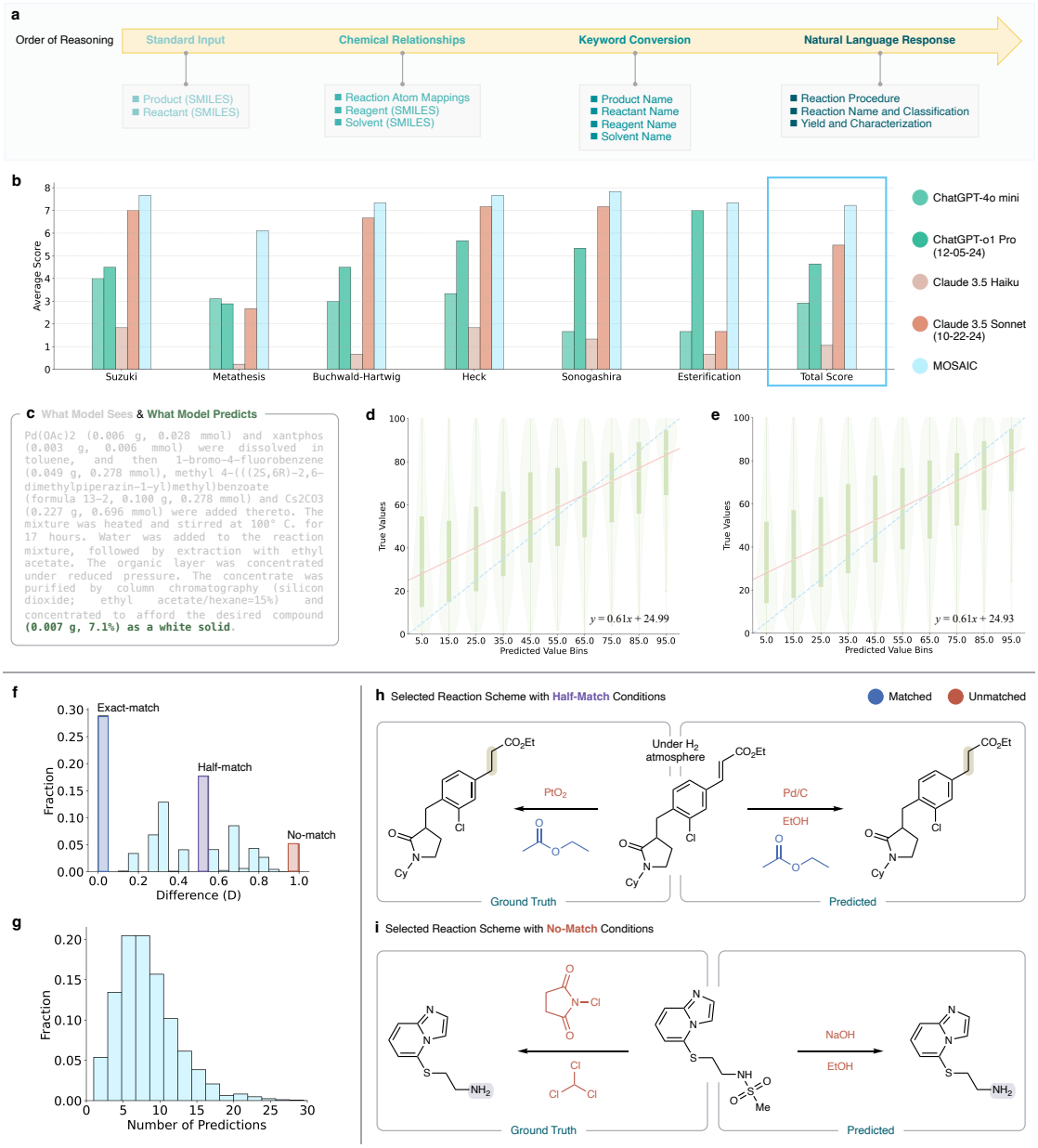

Figure 2: **Prompt design and quantitative metrics**. **a**, prompt design. The prompt is split into four sections. The first part introduces the product and reactant which define the reaction transformation. Reagents, solvents, and atom mappings relevant to the reaction are then completed in the same SMILES format. To transition to full natural language descriptions, the SMILES representations from the previous steps are translated into their chemical names. Lastly, the full reaction procedures along with a possible reaction classification and yield characterization are predicted based on the available information. **b**, based on the same prompt in **a** MOSAIC is compared with other publicly accessible as well as subscription-only LLMs. The result shows superior performance from MOSAIC across all reaction categories and in the averaged final scores. Responses with scores below 5.0 are often not usable for real synthetic practices. **c**, partial paragraph information processing. To obtain yield prediction, the model is provided a partial paragraph before the yield information. The predictions are averaged across beam-searched results. **d**, **e**, comparative yield prediction analysis showing results with reaction classes capped at 20 examples (**d**) and unconstrained (**e**). Each plot combines box and violin plots to display true yield distributions within prediction bins, with median-based fit lines plotted against perfect correlation (dashed). **f**, distribution analysis of prediction accuracy, measured as the percentage overlap between predicted and true reagent/solvent sets. A zero difference indicates an exact match. **g**, frequency distribution of total predictions generated across three expert models. **h**, half-match of conditions (D = 0.5). The model suggests palladium on carbon as a catalyst and ethanol as a solvent, which are typical for alkene hydrogenation reactions. **i**, no overlap with the reference conditions (D = 1.0). For mesyl (methanesulfonyl) deprotection, the model predicts sodium hydroxide in ethanol, which is a common strategy that experienced chemists often consider first. Both examples show that even when predictions differ from reference conditions, the model suggests reasonable alternatives that could lead to viable synthetic outcomes. Et, ethyl; Cy, cyclohexyl.

molecules:

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

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 were conducted using both one-shot and multiple-shots approaches. In the one-shot case, only the first prediction with highest beam score from the closest expert to the query reaction was considered. And in the multiple-shots approach, predictions from up to three experts were aggregated. Additionally, following [24], we also recorded performance for partial matches, defined as cases where at least one molecule in the true set appears in the predicted set ($D < 1$). The summarized results are presented in Tab. 1.

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, while partial matches increase to 45.4% and 51.7%. When leveraging predictions from multiple experts, the performance improves dramatically with exact matches nearly double for reagents to 43.0%, while solvent prediction accuracy rises to 32.8%. Furthermore, the partial match success rate in multiple-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 demonstrate that consulting multiple experts significantly improves prediction accuracy, nearly doubling the exact match rate for reagents and substantially enhancing partial matches for both reagents and solvents.

Pred. Type	Match	Reagent	Solvent	Both
One-shot	Exact	22.4	29.8	12.9
	Partial	45.4	51.7	73.0
Multiple-shots	Exact	43.0	32.8	28.9
	Partial	76.0	55.2	94.8

Table 1: Reagents and Solvents Prediction Results (in % Matches). "Both": solvents and reagents as one set.

Analysis of Model Behavior: In cases where 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 oftern predicted iron as a reagent instead of the tin chloride present in the true set, a valid alternative rather than hallucination. This differentiation reflects the nuanced expertise of the model; among the top 10 experts for such reactions, all focused on nitro-to-amino chemistry, with one expert (ranked sixth) specialized in tin chloride transformations (details in Supplementary Section 5).

Two representative cases illustraing the "half-match" ($D = 0.5$) and "no-match" ($D = 0$) cases are analyzed in figure 2**h**, **i**. Even no-match cases, MOSAIC's predictions often represent plausible alternatives to achieve the desired transformations. This conclusion is further supported by our successful experimental validations despite the inherent variability in synthetic approaches.

Comparison with Other LLMs

To evaluate the capabilities of current language models in addressing chemistry-specific tasks, we conducted comparisons across diverse and important reaction types. Our assessment included 12 reactions across Suzuki coupling, olefin metathesis, Buchwald-Hartwig amination, Heck reactions, Sonogashira coupling, and esterification applied to novel substrates with varying complexities. These reactions were tested against leading language models: ChatGPT-4o mini, Claude 3.5 Haiku, Claude 3.5 Sonnet (10-22-24), and ChatGPT-o1 Pro (12-05-24). Of these models, ChatGPT-4o mini and Claude 3.5 Haiku are publicly accessible, while Claude 3.5 Sonnet and ChatGPT-o1 Pro are subscriptions-based. The performances of the publicly accessible models are particularly relevant to the broader scientific community.

The evaluation framework considers criteria related to the chemical understanding and experimental feasibility. Models were scored on 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 genuine 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 of their chemical reasoning abilities. To ensure reliability and account for response variations, we repeated each prediction three times using identical prompts which contain an example. (Fig. 2**a**). The results, summarized in Fig. 2**b** and also documented in the Supplementary Section 11–16, corroborates MOSAIC's consistent advantage across all reaction categories.

While operating with only 8 billion parameters compared to the likely orders-of-magnitude larger commercial models such as ChatGPT-o1 Pro and Claude 3.5 Sonnet. Prominently, MOSAIC demonstrated superior performance in the obtaining atom mappings. Whereas commercial LLM models consistently fail on this task despite identifying correct reaction classes. This performance gap suggests that targeted fine-tuning and chemistry-specific optimization can overcome raw parameter count advantages in specialized domains. Our experimental validation of MOSAIC's predictions substantiates its practical utility and indicates

that domain expertise can be effectively encoded in smaller, specialized architectures.

De Novo Compound Synthesis Across Reactions of Broad Impact

We carried out extensive experimental validations to evaluate MOSAIC's ability to predict procedures for challenging and practically important chemical transformations. These experiments focused on reactions critical to drug development, advanced materials, and consumer products. By following MOSAIC's procedures, we aimed at determine its reliability and practical utility in guiding the synthesis of novel compounds.

Experimental Design and Setup: For validation, up to three experts employed per compound were employed, using a beam size of 20 with beam groups ranging from 2 to 20 (detailed in Supplementary Section 3). The highest-ranked predictions from each expert are prioritized (Section Experiment Prioritization), yielding predominantly successful outcomes. For novel transformation types shown in Fig. 4 **13** where prediction confidence was low, additional predictions were tested to ensure comprehensive evaluation. All experimental trials are fully documented in Supplementary Section 17 ensuring methodological transparency and reproducibility.

Key Reaction Types and Results: We specifically chose the three Nobel Prize-winning reactions that form the backbone of organic synthesis in particular modern pharmaceutical and materials development [25, 26, 27, 28, 29].

1.**Suzuki Coupling:** A reaction that revolutionized drug synthesis with its precision and safety, enabled access to complex biaryls (Fig. 3 **2a–d**) containing sensitive functional groups that typically pose significant synthetic challenges [30]. MOSAIC's predictions navigated these complexities, underscoring its utility in pharmaceutical development.

2.**Heck Coupling:** MOSAIC demonstrated applicable in Heck coupling reactions (Fig. 3 **3a–d**) including the cases where prior attempts documented in the literature had failed (Fig. 3 **3d**), demonstrating the system's ability to overcome documented limitations [31].This illustrates MOSAIC's potential to address current synthetic bottlenecks.

3.**Olefin Metathesis:** (Fig. 3 **4a**, **4b**) Reaction was equally successful, enabling manipulation of carbon-carbon double bonds, an essential capability for both small molecule synthesis and polymer science [32, 33]. The late-stage modification of the pressure sensitive adhesive 4-acryloyloxy benzophenone (ABP) [34] showcased the versatility of its applications. Moreover, it allowed us to obtain the monomer of a functional material (Fig. 3 **4b**) via ring opening metathesis (ROM) [35, 36, 37].

4.**Buchwald-Hartwig Aminations:** Beyond these mentioned reactions, we validated MOSAIC on transformations relevant to drug development. The Buchwald-Hartwig amination is responsible for forming carbon-nitrogen bonds, found in over 80% of drug molecules [38]. MOSAIC not only successfully predicted conditions for these challenging reactions, (Fig. 3 **1a–c**) but also showed chemical insight by suggesting both palladium-catalyzed Buchwald-Hartwig and copper-catalyzed Goldberg reactions as viable pathways across different substrates. This versatility proved valuable in synthesizing derivatives of important drugs such as the Nortriptyline (antidepressant) and the Fenofibrate (cholesterol-lowering medication).

5.**Sonogashira Coupling:** MOSAIC excelled in predicting conditions for transformations of alkynes, occurring among natural products to functional materials [39], including dual-catalyst systems for Sonogashira coupling reactions (Fig. 3 **5a**), critical for synthesizing materials used in optoelectronic device development [40]. This capability underscores the model's relevance in advanced material applications.

6.**Diaryl Ether Synthesis:** Diaryl ethers are one of the most common and enduring scaffolds in medicinal chemistry and agrochemicals applications [41]. Among this category, MOSAIC provided multiple viable pathways for synthesizing estrone derivatives (Fig. 3 **6a**, **6b**), effectively adapting established methods for the modification of complex bioactive molecules.

In addition to catalytic reactions, we investigated prevalent reactions in organic synthesis where controlling selectivity and reactivity continues to be a key challenge. MOSAIC's predictions guided experiments across several important reaction classes, showcasing its applicability in complex, chemoselective, and stereoselective transformations.

1.**Chemoselective and Controlled Oxidation**: MOSAIC provided conditions for the controlled oxidation of a derivative of Pentaerythritol, a versatile building block for explosives, plastics, paints, and appliances, to its corresponding aldehyde (Fig. 3 **7a**). Notably, this transformation avoided the use of toxic metal like chromium [42], underscoring the model's utility in providing safer and more sustainable synthetic pathways.

2.**Site- and Stereoselective Transformations**:
2.1. Conjugate Addition to Chiral Pool Materials: MOSAIC predicted successful conjugate addition to (*S*)-carvone (Fig. 3 **7b**), a chiral monoterpene frequently used in stereospecific natural product synthesis. This result demonstrates the model's ability to manage site-selective transformations in complex chiral substrates.

2.2. Prenylation of Sclareolide: The sesquiterpene lactone natural product Sclareolide underwent a prenylation reaction (Fig. 3 **7e**), an important reaction used in post-translational modification to increase the structural diversity and bioactivity of peptides and proteins [43]. This showcases MOSAIC's ability to predict conditions for selective modifications of biologically relevant molecules.

2.3. Olefination of Carbonyl Compounds: MOSAIC successfully predicted Horner-Wadsworth-Emmons reaction conditions for stereoselective olefination of *L*-

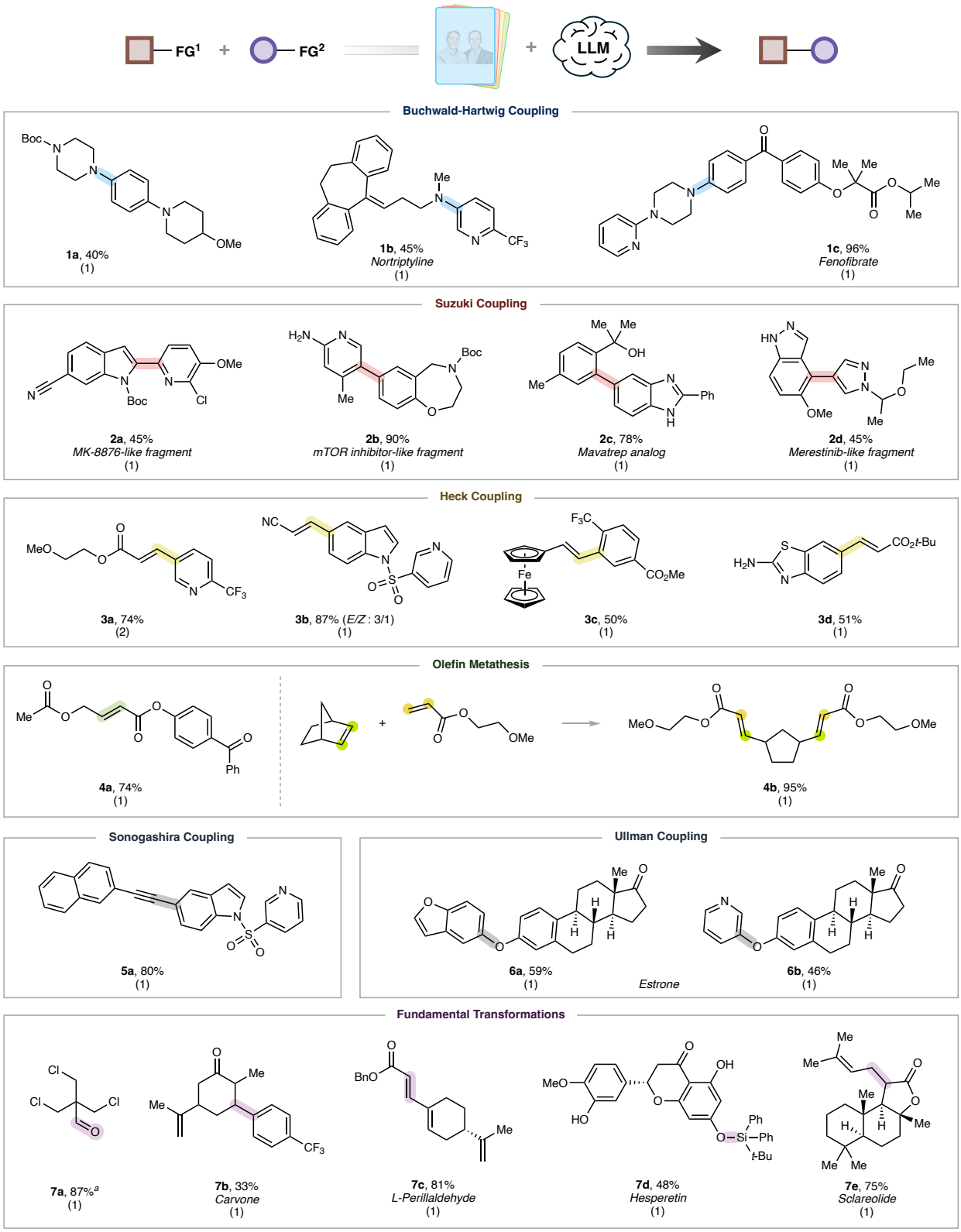

Figure 3: **MOSAIC-guided diverse de novo compound synthesis.** All yields are isolated unless otherwise indicated. [a] NMR yields. Number of trials are in parenthesis. See Supplementary Information for experimental details. Boc, tert-butyloxycarbonyl; Me, methyl; *t*-Bu, tert-butyl; Ph, phenyl; Bn, benzyl.

Perillaldehyde (Fig. 3 **7c**), a monoterpenoid volatile oil derived from perilla herb, widely used in the flavor and perfumery industries [44].

3.**Site-Specific Functional Group Manipulation:** Selective silyl protection of Hesperetin(Fig. 3 **7d**), a naturally occurring antioxidant, and anti-inflammatory agent, was successfully achieved using MOSAIC's procedures, highlighting it's ability to orchestrate complex functional group reactivity in intricate molecular scaffolds, providing viable and selective synthetic procedures.

These results validate the model's versatility for a widespread transformations including natural product synthesis and derivatization.

Translational Applications

The practical impact of MOSAIC spans diverse chemical industries, from pharmaceuticals to advanced materials, catalysis, agriculture, and cosmetics. These translational applications underscore its versatility and potential to drive innovation in both established and emerging fields.

Pharmaceutical Development: Our validation studies encompassed both the creation of new drug-like molecules (Fig. 4 **8a**) and strategic modifications of existing drugs (Fig. 4 **8b**–**d**). These capabilities prove critical for enhancing drug safety, efficacy, and pharmacokinetic properties [45]. Through precise predictions of functionalization conditions and novel scaffold construction, the system enables rapid prototyping of therapeutic candidates.

Catalysis: In the realm of catalyst design, we achieved synthesis of the specialized ligand **9a** (Fig. 4) – a bipyridine ligand, essential for diverse catalytic applications including industrial chemical processes [46]. Notablty, predicted amination conditions afforeded new analogs of photocatalyst (Fig. 4 **9b**, **9c**), compounds that harness light energy to drive chemical reactions. This advancement in sustainable chemistry holds significant promise for reducing environmental impact [47].

Advanced Materials: Through systematic prediction of reaction conditions, novel conjugated compounds (Fig. 4 **10a**, **10b**) emerged as promising candidates for electronic devices, including organic semiconductors and light-emitting diodes. While larger molecules demanded more extensive computational searches, the resulting predictions proved actionable [48], highlighting the system's potential in functional materials science.

Agrochemicals: The development of new agrochemical variants, particularly derivatives of pyrabactin (Fig. 4 **11a**–**c**) [49], demonstrated promising results. These advances pave the way for more effective and environmentally conscious agricultural practices.

Cosmetics and Fragrance Development: Our investigations yielded new fragrance analogs (Fig. 4 **12a**, **12b**) derived from hedione. In addition, we have obtained citronellylretinoate (Fig. 4 **12c**), a variant of anti-aging chemeical retinylretinoate [50]. The successful prediction of synthesis protocols for these complex cosmetic ingredients underscores the system's value in consumer-focused applications.

Methodology Discovery: The realm of translational applications expands with the development of new methods as a result of the new tools and strategies facilitating chemical space exploration, importance of which can even be traced back to several Nobel Prizes in recent years including the ones mentioned earlier [51]. However, the development of new chemical transformations relies on a thorough understanding of reaction mechanisms, fundamental reactivity trends, and the underlying principles that govern them, requires trial and error processes for fine adjustments of the reaction conditions. Gratifyingly, MOSAIC's capability of processing the collective knowledge enables to guide new reaction discovery, exemplified by new heterocycle ring synthesis protocols (Fig. 4 **13a**) [52], allowing access to under-explored bioisosteric analogs of indoles [5]. This methodological versatility expands the synthetic toolkit while enabling exploration of novel chemical space

These diverse applications highlight the system's potential to accelerate innovation across the chemical sciences, from therapeutic development to sustainable materials and consumer products. The demonstrated ability to predict conditions for a wide range of practically significant transformations indicates readiness for deployment in both research and industrial settings. Furthermore, these results validate our domain specialization approach, offering distinct advantages over existing LLMs that struggle with standard chemical notations and lack robust confidence measurements in their predictions.

Discussion

Model Limitations and Future Directions

While MOSAIC demonstrates remarkable capabilities in predicting reliable synthetic procedures, it operates within certain boundaries that reflect the current 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.

For specialized applications, MOSAIC's general-purpose architecture necessarily trades some precision for breadth. For instance, while it can predict yields for specific Buchwald-Hartwig reactions between 4-methylaniline and aryl halides, it does not match the precision of bespoke models [4] that are optimized for this task using curated datasets [53]. However, this trade-off enables MOSAIC to address a vastly broader range of chemical challenges, making it versatile for real-world applications.

The current implementation inherits constraints from standard LLM tokenization strategies. While these tokens effectively capture chemical knowledge [54], they could

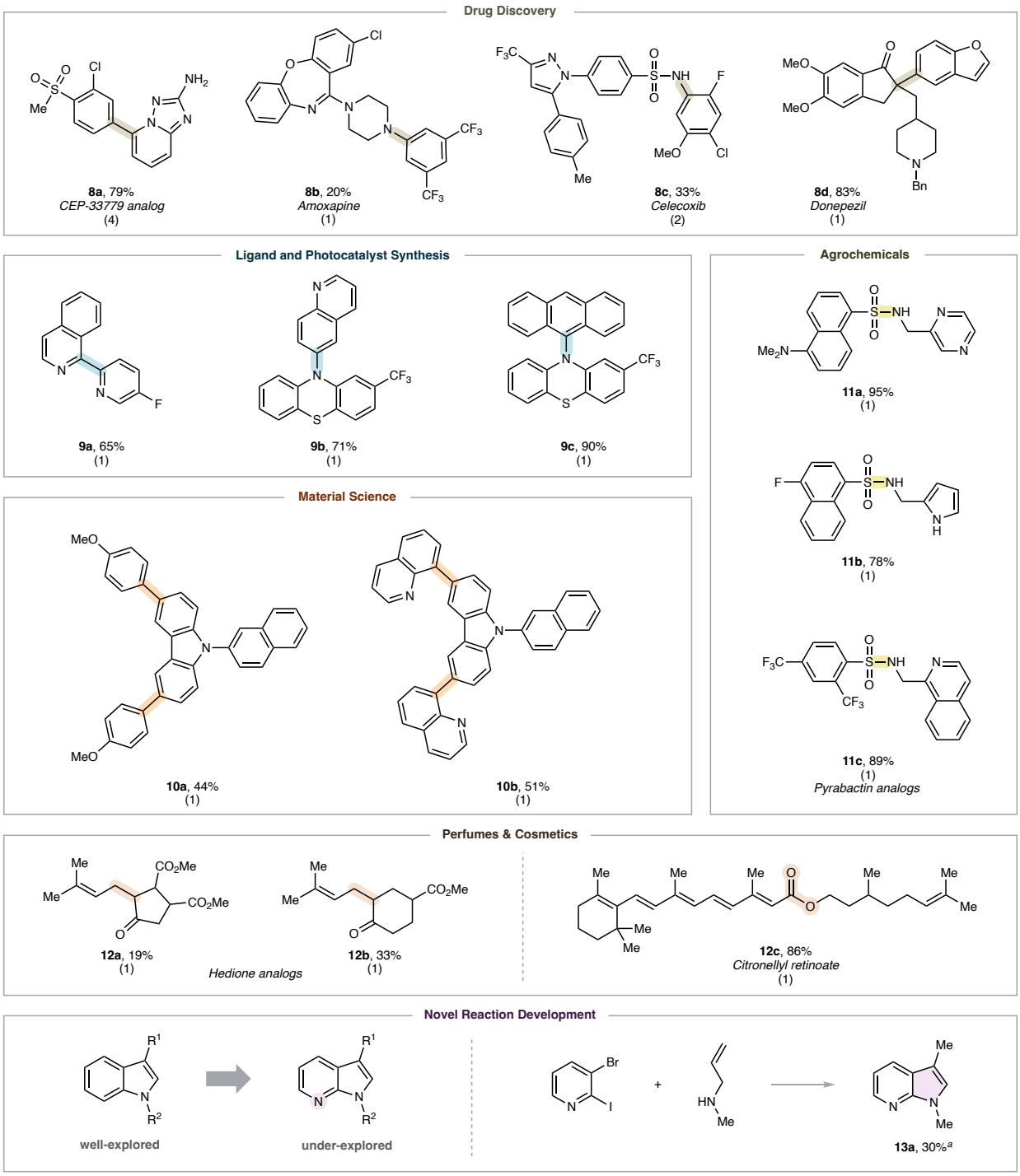

Figure 4: **New compounds synthesized for translational applications.** All yields are isolated unless otherwise indicated. [a] NMR yields. Number of trials are in parenthesis. See Supplementary Information for experimental details. Me, methyl; Bn, benzyl.

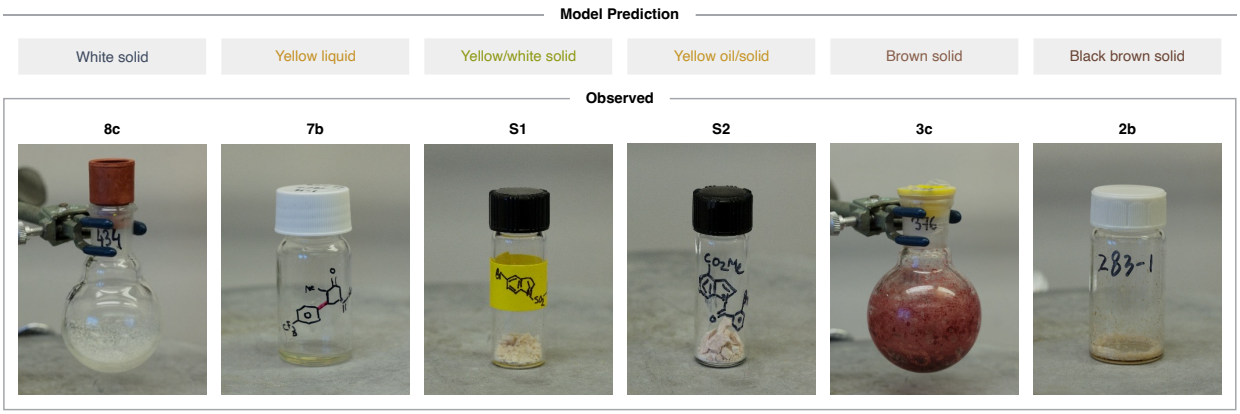

Figure 5: **Examples of successful applications of MOSAIC**. By following the provided synthesis and workup procedures precisely, the separated compounds (Figures 3, 4 and SI) were compared to the model-predicted products.

struggle with complex molecular structures, particularly in translating between SMILES notation and compound names for large molecules with multiple heterocyclic rings. Interestingly, even when nomenclature details are imperfect, the model maintains accurate recognition of functional groups and reaction patterns, demonstrating robust chemical understanding (Supplementary Section 17).

These observations point to exciting opportunities for future development. Implementation of chemistry-specific tokenization approaches, such as mixed atom-character encoding [55] or multi-modal representations with explicit connectivity information [56], could further enhance MOSAIC's performance. Moreover, while our implementation uses Llama-3.1-8B-instruct as the base model, the framework's architecture is model-agnostic and could seamlessly incorporate larger models such as Llama-3.1-70B or 405B [15], or more recent versions such as Llama-3.3 series, which demonstrate superior performance in general language tasks. This flexibility, combined with MOSAIC's scalable architecture, ensures that it can readily incorporate future advances in both language modeling and chemical representation, further narrowing the gap between computational prediction and experimental outcomes.

Delicate Control of Reaction Metrics

One of MOSAIC's key strengths lies in its ability to predict detailed procedural steps that are critical for synthetic success. Beyond basic reaction conditions, the system provides precise guidance on factors such as reagent addition order, temperature control, and workup procedures—elements traditionally refined through years of laboratory experience.

Here we discuss several examples to show how these procedures become decisive:

1.**Oxidative Transformations:** When using mCPBA, initial addition at -20°C followed by controlled warming prevents side reactions such as epoxide ring-opening side reactions [57].

2.**Nozaki-Hiyama-Kishi Coupling:** Premature addition of oxidants, before complete formation of organometallic intermediates, leads to substrate degradation, while slow addition over 30–60 minutes ensures selective carbonyl addition [58].

3.**S_N2 Reactions:** Trace catalytic additives like iodide salts (5–10 mol%) can dramatically increase yields from <10% to >90% through halide exchange catalysis [59].

4.**Sonogashira Coupling:** Copper(I) iodide facilitates transmetallation reaction with palladium catalyst that enables reactions at lower temperatures [40].

5.**Photoredox Transformations:** Precise setup protocols, such as degassing of the reaction mixture (e.g., freeze-pump-thaw cycles and nitrogen bubbling) prevents oxygen quenching of excited states, while optimal light source positioning (e.g. 3–5 cm from the reaction vessel) ensures sufficient photon flux without causing localized heating [60].

6.**Ring-Closing Metathesis:** Adding 1–5 mol% 1,4-benzoquinone prevents catalyst decomposition [61].

7.**C–H Functionalization:** Addition of catalytic amount of pivalic acid (30 mol%) in palladium-catalyzed C–H arylation of arenes generates a highly active catalyst, while minimizes the formation of undesired byproducts [62].

8.**Workup Procedures:** Proper uses of specific buffer solutions and saturated ammonium chloride solution prevent decomposition of sensitive products [63].

Unlike bespoke models that often omit such details, MOSAIC systematically captures these nuances, allowing for more comprehensive and accurate predictions. Mastering these nuances of chemical synthesis traditionally demands years of laboratory experience, yet in an era where chemical knowledge expands at an unprecedented rate, even decades of expertise cannot keep pace with the expanding frontiers of synthetic possibility. This intrinsic limitation in knowledge acquisition and synthesis planning will lead

researchers to naturally gravitate toward computational frameworks like MOSAIC which systematically processes, integrates, and applies the collective intelligence.

Information Overload to Actionable Knowledge

The challenge of identifying viable reaction procedures is also evidenced by the paradox in modern chemical search systems: the wealth of available information can be as problematic as its absence. A typical structure-based search in chemical databases such as SciFinder can return hundreds of thousands of related procedures, creating what appears to be a comprehensive resource but effectively presents a time-consuming challenge to determine which procedure to implement. For this reason, chemists may ignore this abundance of data and instead determine the optimal conditions through iterative optimizations.

MOSAIC addresses this information overload through the domain-expert design. Rather than using all possible information, only the most relevant is used followed by the decision from the LLM itself. This approach effectively distills vast amounts of reaction data into actionable procedures that guide experimental outcomes.

Confidence Estimate

One of the most significant advancements in the application of large language models to chemistry is the ability to provide reliable confidence estimates. In MOSAIC, confidence metrics are inherently data-driven, derived from the distance between a query reaction and the centroid of the domain expert responsible for the prediction. This distance-based confidence measure offers intuitive interpretation: predictions from experts with centroids closer to the query reaction carry higher confidence.

Our analysis, detailed in Appendix figures (Fig. 6 and 7), reveals distinct confidence thresholds that correlate with prediction reliability:

1.Distances < 50: High confidence predictions with strong structural and mechanistic resemblance, sharing similar transformation patterns and maintaining close similarity in both reactant and product structures.

2.Distances of 100–200: Moderate confidence predictions, retaining core transformation patterns while showing greater variation in substrate and product structures.

3.Distances > 200: Predictions falling within the same broad reaction category but often involving different reactive groups or reaction conditions.

Importantly, even predictions with lower confidence scores can provide valuable insights. For instance, when querying a Buchwald-Hartwig coupling reaction of aryl bromides at a distance of 300, analogous couplings with aryl iodides were identified. Such matches can serve as useful inspirations or suggest alternative synthetic approaches.

Extended Utilities

Beyond its current applications, the Reaction Similarity Framework Platform (RSFP) serves as a robust and versatile tool for chemical knowledge retrieval and analysis. By referencing analogous reactions, RSFP facilitates data-driven assessments of critical parameters such as reaction scalability, green chemistry metrics, cost-effectiveness, and synthetic feasibility. These assessments can be used to narrow down massive synthetic possibilities such as those generated by retrosynthesis modeling tools [64, 65, 66, 67, 68] before MOSAIC provides detailed experimental procedures for the most promising pathways.

The scalability of the framework suggests significant potential from integration with public repositories [69, 70] and comprehensive databases such as Reaxys and SciFinder, which contain over 100 million document records, nearly 100 times of our current training data. This could enable the stronger generalization and recognition of emerging synthetic methods as domain knowledge, including specialized approaches in photochemistry and electrochemistry. This attribute ensures that MOSAIC remain aligned with the rapidly evolving landscape of chemical synthesis, enabling it to address increasingly complex synthetic challenges.

Conclusion

Our development of MOSAIC embodies the principle that methods leveraging the computational search tend to scale effectively with increasing amount 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 becomes available. The search mechanism through FAISS enables efficient navigation of this chemical space, allowing the system to quickly identify the most relevant expert models for any given query. This architecture allows us to grow the number of expert models as new reaction classes emerge while maintaining an efficient search through the centroid-based approach. Importantly, this approach avoids the limitations of strict definitions of reaction types, instead allowing the system to discover and utilize similarities across transformation patterns directly from the Voronoi cells.

Chemists have already adapted to the use of many changes to the way literature is accessed from the physical book to online depositories, and the advent of large language models offers to provide the next transition. We envision MOSAIC to function as an essential navigational tool in modern chemical synthesis. The integration of large language models with comprehensive reaction databases creates a powerful in silico platform that enables chemists to systematically obtain reaction procedures and identify viable synthetic routes with unprecedented speed and precision. What once required extensive human labor for each reaction to determine suitable conditions can now be ac-

complished within minutes, dramatically accelerating the pace of discovery. While seasoned chemists may not rely on such tools for routine transformations, these computational frameworks are becoming increasingly indispensable in contemporary laboratory practice, accelerating scientific discoveries across organic synthesis, pharmaceutical innovation, advanced materials science, and agrochemical research. Chemistry remains an empirical science, where novel methodologies often emerge through meticulous mechanistic investigation and serendipitous discovery. By integrating this empirical insights with advanced modeling approach, we bridge the intuition of chemical synthesis with data-driven inspirations. This approach significantly reduces the time, resources, and environmental impact associated with reaction development and optimization, while simultaneously expanding the boundaries of synthetic possibilities. This synergy fosters a powerful cycle for addressing complex synthetic challenges, empowering breakthroughs in both academic and industrial settings.

Prompt Design

We developed a structured prompt template inspired by the Alpaca framework's input-response format [71], specifically adapted for chemical contexts. The template logically organizes chemical information to enable reaction prediction and procedure generation. Our implementation uses the structure described in Fig. 2**a**. The prompt structure leverages the inherent auto-regressive prediction capabilities of transformer-decoder architectures [72] 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 via reaction SMILES strings, incorporating both reagents and solvents. The model then derives specific reagents and solvent SMILES from the complete reaction mapping. 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 based on the aggregated 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, where each expert e provides M_e predictions (M_e vary by expert), we establish a ranking function $R(e, p)$ where $e \in [1, N]$ 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 where 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.

Safety Guidelines

All chemical procedures produced by MOSAIC must only be carried out by individuals with appropriate safety training and within 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 established safety protocols.

Acknowledgments

The authors acknowledges 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 award BES-ERCAP0030682, BES-ERCAP0024372, and BES-ERCAP0027329 and partial support for the development of the MOSAIC software from the National Science Foundation Engines Development Award: Advancing Quantum Technologies (CT) under Award Number 2302908.

H.L. gratefully acknowledges helpful discussions with Jordan E. Crivelli-Decker, and Benjamin J. Shields et al. at SandboxAQ. H.L. thanks Inomi for contributing hand-drawn art works related to Fig.1d.

Code Availability

The code and annotated notebooks are available at https://github.com/haoteli/MOSAIC.

Data Availability

The pistachio database (version number 2024Q1) can be accessed at https://www.nextmovesoftware.com/pistachio. Llama-3.1-8B-Instruct model weights are accessed at https://www.llama.com/llama-downloads

Author contributions

H.L conceived the method, developed the theoretical framework, implemented the computational applications and

wrote the initial manuscript. S.S designed and performed experimental validations, defined and examined the application scope, and made crucial contributions 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 on the method. T.R.N and V.S.B acquired funding. T.R.N, V.S.B, R.H.C supervised this research and revised the manuscript for publication. H.L and S.S jointly prepared the final manuscript.

References

[1] Andrew D White. The future of chemistry is language. *Nature Reviews Chemistry*, 7(7):457–458, 2023.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Hanyu Gao, Thomas J Struble, Connor W Coley, Yuran Wang, William H Green, and Klavs F Jensen. Using machine learning to predict suitable conditions for organic reactions. *ACS central science*, 4(11):1465–1476, 2018.

[4] Shu-Wen Li, Li-Cheng Xu, Cheng Zhang, Shuo-Qing Zhang, and Xin Hong. Reaction performance prediction with an extrapolative and interpretable graph model based on chemical knowledge. *Nature Communications*, 14(1):3569, 2023.

[5] Manajit Das, Ankit Ghosh, and Raghavan B Sunoj. Advances in machine learning with chemical language models in molecular property and reaction outcome predictions. *Journal of Computational Chemistry*, 45(14):1160–1176, 2024.

[6] Alain C Vaucher, Philippe Schwaller, Joppe Geluykens, Vishnu H Nair, Anna Iuliano, and Teodoro Laino. Inferring experimental procedures from text-based representations of chemical reactions. *Nature communications*, 12(1):2573, 2021.

[7] Mayk Caldas Ramos, Christopher J Collison, and Andrew D White. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 2025.

[8] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[9] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11, 2024.

[10] Chonghuan Zhang, Qianghua Lin, Biwei Zhu, Haopeng Yang, Xiao Lian, Hao Deng, Jiajun Zheng, and Kuangbiao Liao. Synask: unleashing the power of large language models in organic synthesis. *Chemical Science*, 16(1):43–56, 2025.

[11] Matthew Renze and Erhan Guven. The effect of sampling temperature on problem solving in large language models. *arXiv preprint arXiv:2402.05201*, 2024.

[12] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[13] Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, et al. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature Communications*, 15(1):10160, 2024.

[14] Connor W Coley, Dale A Thomas III, Justin AM Lummiss, Jonathan N Jaworski, Christopher P Breen, Victor Schultz, Travis Hart, Joshua S Fishman, Luke Rogers, Hanyu Gao, et al. A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science*, 365(6453):eaax1566, 2019.

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[17] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

[18] NextMove Software. Pistachio, 2024. Accessed: June 2024.

[19] Daniel Probst and Jean-Louis Reymond. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics*, 12(1):12, 2020.

[20] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

[21] Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pages 1–14, 2021.

[22] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

[23] Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM computing surveys (CSUR)*, 38(2):6–es, 2006.

[24] Mikhail Andronov, Varvara Voinarovska, Natalia Andronova, Michael Wand, Djork-Arné Clevert, and Jürgen Schmidhuber. Reagent prediction with a molecular transformer improves reaction data quality. *Chemical Science*, 14(12):3235–3246, 2023.

[25] Javier Magano and Joshua R Dunetz. *Transition metal-catalyzed couplings in process chemistry: case studies from the pharmaceutical industry*. John Wiley & Sons, 2013.

[26] Javier Magano and Joshua R Dunetz. Large-scale applications of transition metal-catalyzed couplings for the synthesis of pharmaceuticals. *Chemical Reviews*, 111(3):2177–2250, 2011.

[27] Xiao-Feng Wu, Pazhamalai Anbarasan, Helfried Neumann, and Matthias Beller. From noble metal to nobel prize: palladium-catalyzed coupling reactions as key methods in organic synthesis. *Angewandte Chemie International Edition*, 48(49):9047–9050, 2010.

[28] Carin CC Johansson Seechurn, Matthew O Kitching, Thomas J Colacot, and Victor Snieckus. Palladium-catalyzed cross-coupling: a historical contextual perspective to the 2010 nobel prize. *Angewandte Chemie International Edition*, 51(21):5062–5085, 2012.

[29] Paula Ruiz-Castillo and Stephen L Buchwald. Applications of palladium-catalyzed c–n cross-coupling reactions. *Chemical reviews*, 116(19):12564–12649, 2016.

[30] Barbara Zdrazil and Rajarshi Guha. The rise and fall of a scaffold: a trend analysis of scaffolds in the medicinal chemistry literature. *Journal of Medicinal Chemistry*, 61(11):4688–4703, 2017.

[31] Wenxin Lu, Haote Li, Jan Paul Menzel, Abbigayle E. Cuomo, Andrea M. Nikolic, H. Ray Kelly, Yu Shee, Sanil Sreekumar, Frederic Buono, Jinhua J. Song, and et al. Enhanced ligand discovery through generative ai and latent-space exploration: Application to the mizoroki-heck reaction. *ChemRxiv*, 2024.

[32] Amir H Hoveyda, Steven J Malcolmson, Simon J Meek, and Adil R Zhugralin. Catalytic enantioselective olefin metathesis in natural product synthesis. chiral metal-based complexes that deliver high enantioselectivity and more. *Angewandte Chemie International Edition*, 49(1):34–44, 2010.

[33] Fern Sinclair, Mohammed Alkattan, Joëlle Prunet, and Michael P Shaver. Olefin cross metathesis and ring-closing metathesis in polymer chemistry. *Polymer Chemistry*, 8(22):3385–3398, 2017.

[34] Zbigniew Czech and Agnieszka Butwin. Development of photoreactive uv-crosslinkable solvent-free acrylic pressure-sensitive adhesives coated at room temperature and used for removable and repositionable self-adhesive materials. *Polish Journal of Chemical Technology*, 13(1):31–34, 2011.

[35] G Alex Cortez, Carl A Baxter, Richard R Schrock, and Amir H Hoveyda. Comparison of ru-and mo-based chiral olefin metathesis catalysts. complementarity in asymmetric ring-opening/cross-metathesis reactions of oxa-and azabicycles. *Organic Letters*, 9(15):2871–2874, 2007.

[36] Wolfram Stüer, Justin Wolf, Helmut Werner, Peter Schwab, and Michael Schulz. Carbynehydridoruthenium complexes as catalysts for the selective, ring-opening metathesis of cyclopentene with methyl acrylate. *Angewandte Chemie International Edition*, 37(24):3421–3423, 1998.

[37] Tae-Lim Choi, Choon Woo Lee, Arnab K Chatterjee, and Robert H Grubbs. Olefin metathesis involving ruthenium enoic carbene complexes. *Journal of the American Chemical Society*, 123(42):10417–10418, 2001.

[38] Ruth Dorel, Christian P Grugel, and Alexander M Haydl. The buchwald–hartwig amination after 25 years. *Angewandte Chemie International Edition*, 58(48):17118–17129, 2019.

[39] Rolf Gleiter and Daniel B. Werz. Alkynes between main group elements: from dumbbells via rods to squares and tubes. *Chemical Reviews*, 110(7):4447–4488, 2010.

[40] Rafael Chinchilla and Carmen Nájera. The sonogashira reaction: a booming methodology in synthetic organic chemistry. *Chemical reviews*, 107(3):874–922, 2007.

[41] Tao Chen, Hao Xiong, Jing-Fang Yang, Xiao-Lei Zhu, Ren-Yu Qu, and Guang-Fu Yang. Diaryl ether: a privileged scaffold for drug and agrochemical discovery. *Journal of Agricultural and Food Chemistry*, 68(37):9839–9877, 2020.

[42] Anthony J Mancuso, Shui-Lung Huang, and Daniel Swern. Oxidation of long-chain and related alcohols to carbonyls by dimethyl sulfoxide" activated" by oxalyl chloride. *The Journal of Organic Chemistry*, 43(12):2480–2482, 1978.

[43] Charuta C Palsuledesai and Mark D Distefano. Protein prenylation: enzymes, therapeutics, and biotechnology applications. *ACS chemical biology*, 10(1):51–62, 2015.

[44] Dávid Roman, Maria Sauer, and Christine Beemelmanns. Applications of the horner–wadsworth–emmons olefination in modern natural product synthesis. *Synthesis*, 53(16):2713–2739, 2021.

[45] Nathan J Castellino, Andrew P Montgomery, Jonathan J Danon, and Michael Kassiou. Late-stage

functionalization for improving drug-like molecular properties. *Chemical reviews*, 123(13):8127–8153, 2023.

[46] Christian Kaes, Alexander Katz, and Mir Wais Hosseini. Bipyridine: the most widely used ligand. a review of molecules comprising at least two 2, 2 '-bipyridine units. *Chemical reviews*, 100(10):3553–3590, 2000.

[47] Nicolas J Treat, Hazel Sprafke, John W Kramer, Paul G Clark, Bryan E Barton, Javier Read de Alaniz, Brett P Fors, and Craig J Hawker. Metal-free atom transfer radical polymerization. *Journal of the American Chemical Society*, 136(45):16096–16101, 2014.

[48] Pierre M Beaujuge and John R Reynolds. Color control in π-conjugated organic polymers for use in electrochromic devices. *Chemical reviews*, 110(1):268–320, 2010.

[49] Sang-youl Park, Pauline Fung, Noriyuki Nishimura, Davin R Jensen, Hiroaki Fujii, Yang Zhao, Shelley Lumba, Julia Santiago, Americo Rodrigues, F Chow Tsz-fung, et al. Abscisic acid inhibits pp2cs via the pyr/pyl family of aba-binding start proteins. *Science (New York, NY)*, 324(5930):1068, 2009.

[50] Hyojung Kim, Bora Kim, Hyuk Kim, Soojong Um, Joodong Lee, Heechang Ryoo, and Hyungil Jung. Synthesis and in vitro biological activity of retinyl retinoate, a novel hybrid retinoid derivative. *Bioorganic & medicinal chemistry*, 16(12):6387–6393, 2008.

[51] Kevin R Campos, Paul J Coleman, Juan C Alvarez, Spencer D Dreher, Robert M Garbaccio, Nicholas K Terrett, Richard D Tillyer, Matthew D Truppo, and Emma R Parmee. The importance of synthetic chemistry in the pharmaceutical industry. *Science*, 363(6424):eaat0805, 2019.

[52] Thomas Jensen, Henrik Pedersen, Benny Bang-Andersen, Robert Madsen, and Morten Jørgensen. Palladium-catalyzed aryl amination–heck cyclization cascade: A one-flask approach to 3-substituted indoles. *Angewandte Chemie International Edition*, 47(5):888–890, 2008.

[53] Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. Predicting reaction performance in c–n cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018.

[54] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[55] Haote Li, Yu Shee, Brandon Allen, Federica Maschietto, Anton Morgunov, and Victor Batista. Kernel-elastic autoencoder for molecular design. *PNAS nexus*, 3(4):pgae168, 2024.

[56] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

[57] Michael Palucki, Paul J Pospisil, Wei Zhang, and Eric N Jacobsen. Highly enantioselective, low-temperature epoxidation of styrene. *Journal of the American Chemical Society*, 116(20):9333–9334, 1994.

[58] Yang Gao, David E Hill, Wei Hao, Brendon J McNicholas, Julien C Vantourout, Ryan G Hadt, Sarah E Reisman, Donna G Blackmond, and Phil S Baran. Electrochemical nozaki–hiyama–kishi coupling: scope, applications, and mechanism. *Journal of the American Chemical Society*, 143(25):9478–9488, 2021.

[59] Hans Finkelstein. Darstellung organischer jodide aus den entsprechenden bromiden und chloriden. *Berichte der deutschen chemischen Gesellschaft*, 43(2):1528–1532, 1910.

[60] Kazuki Miyazawa, Takashi Koike, and Munetaka Akita. Hydroaminomethylation of olefins with aminomethyltrifluoroborate by photoredox catalysis. *Advanced Synthesis & Catalysis*, 356(13):2749–2755, 2014.

[61] Soon Hyeok Hong, Daniel P Sanders, Choon Woo Lee, and Robert H Grubbs. Prevention of undesirable isomerization during olefin metathesis. *Journal of the American Chemical Society*, 127(49):17160–17161, 2005.

[62] Marc Lafrance and Keith Fagnou. Palladium-catalyzed benzene arylation: incorporation of catalytic pivalic acid as a proton shuttle and a key element in catalyst design. *Journal of the American Chemical Society*, 128(51):16496–16497, 2006.

[63] Jerry R. Mohrig, David Alberg, Gretchen Hofmeister, Paul F. Schatz, and Christina Noring Hammond. *Laboratory Techniques in Organic Chemistry*. W. H. Freeman and Company, New York, 4th edition, 2014.

[64] Yu Shee, Haote Li, Pengpeng Zhang, Andrea M Nikolic, Wenxin Lu, H Ray Kelly, Vidhyadhar Manee, Sanil Sreekumar, Frederic G Buono, Jinhua J Song, et al. Site-specific template generative approach for retrosynthetic planning. *Nature Communications*, 15(1):7818, 2024.

[65] Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.

[66] Yu Wang, Chao Pang, Yuzhe Wang, Junru Jin, Jingjie Zhang, Xiangxiang Zeng, Ran Su, Quan Zou, and Leyi Wei. Retrosynthesis prediction with an interpretable deep-learning framework based on molecular assembly tasks. *Nature Communications*, 14(1):6155, 2023.

[67] Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp

transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):5575, 2020.

[68] Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020.

[69] Steven M Kearnes, Michael R Maser, Michael Wleklinski, Anton Kast, Abigail G Doyle, Spencer D Dreher, Joel M Hawkins, Klavs F Jensen, and Connor W Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021.

[70] Samuel Genheden and Esben Bjerrum. Paroutes: towards a framework for benchmarking retrosynthesis route predictions. *Digital Discovery*, 1(4):527–539, 2022.

[71] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[72] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

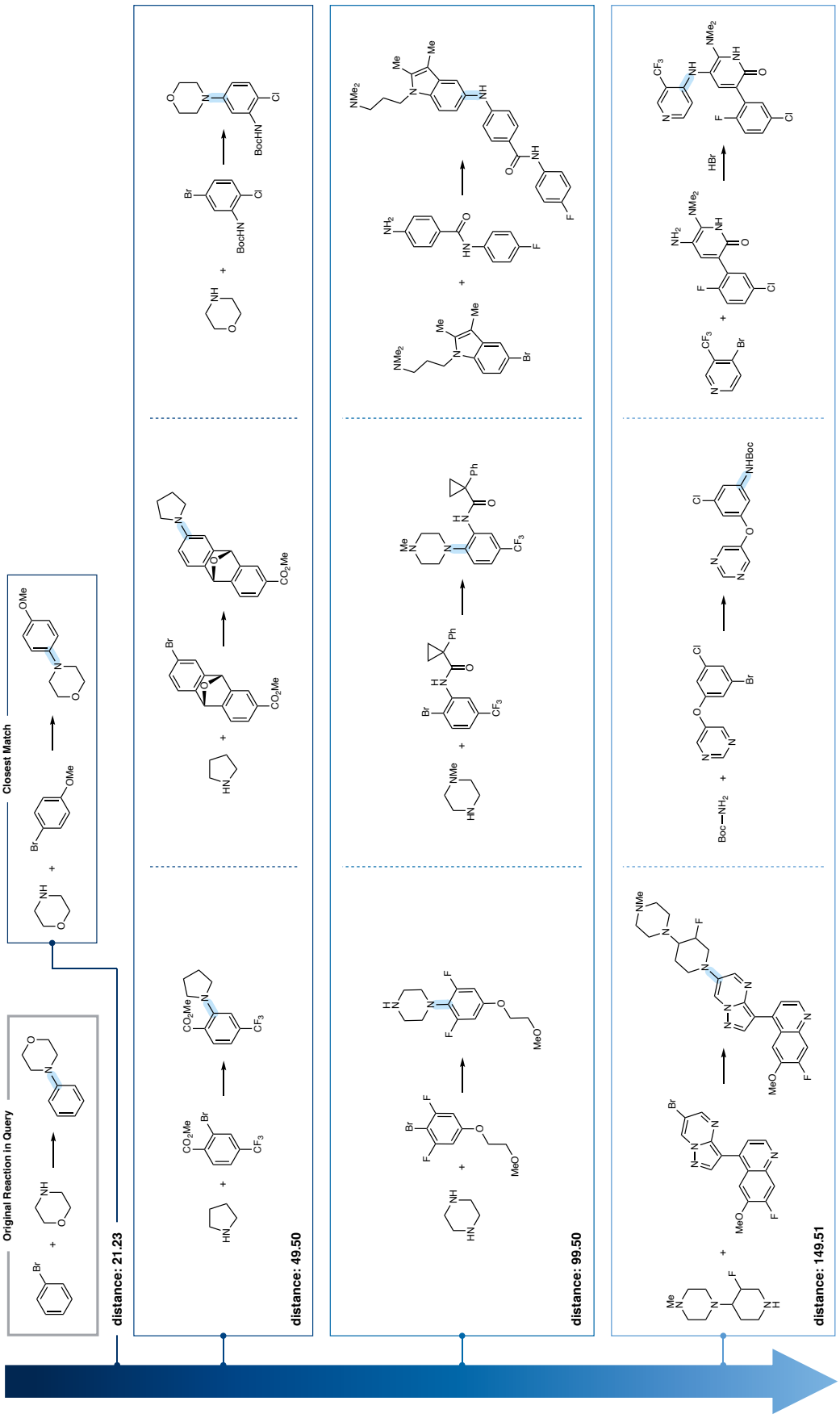

Figure 6: Referencing reactions at smaller distances (higher confidence)

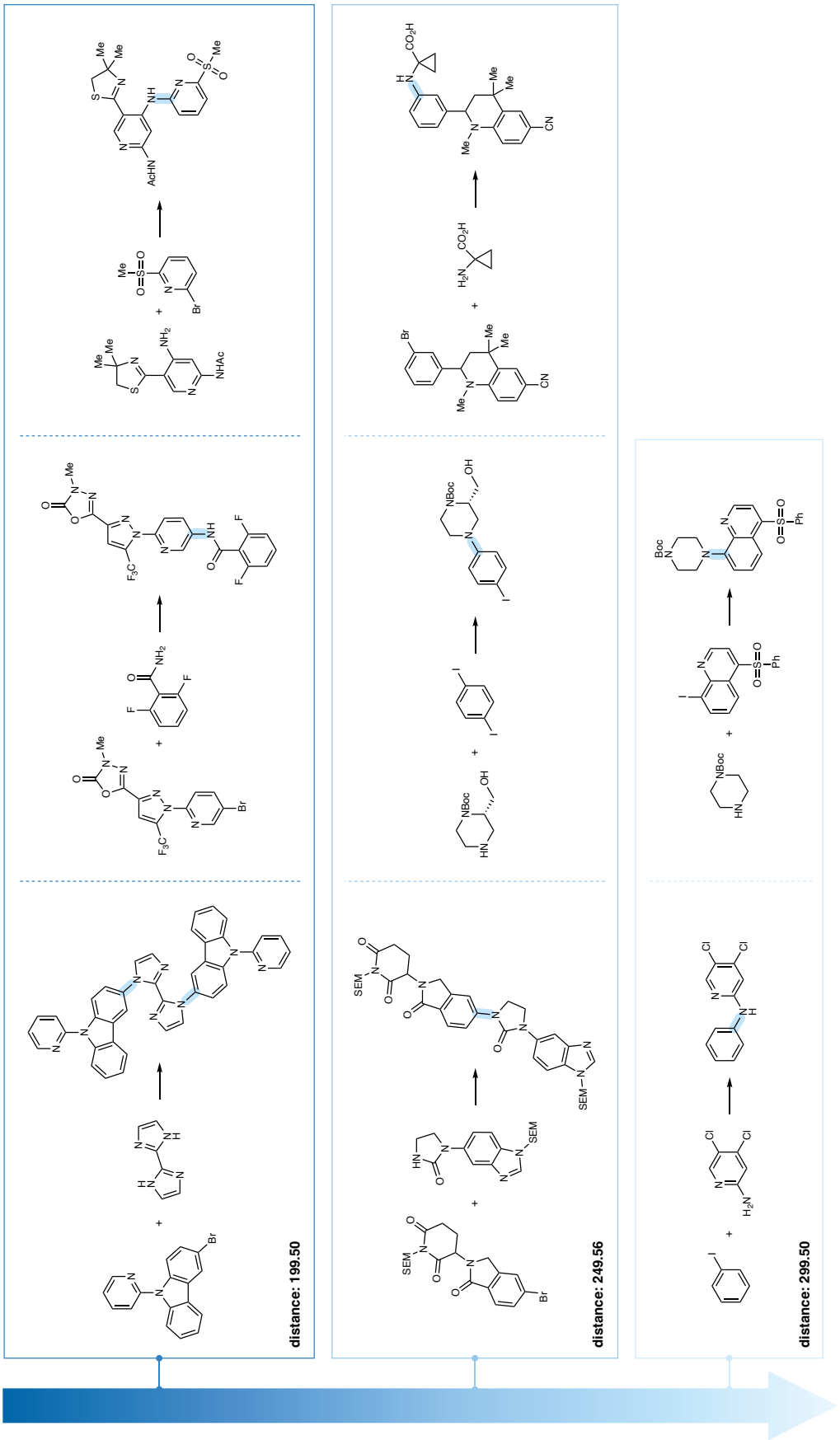

Figure 7: Referencing Reactions at larger distances (lower confidence)