REVEALING CHEMICAL REASONING IN LLMS THROUGH SEARCH ON COMPLEX PLANNING TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have been the focal point of enormous development in artificial intelligence over the past half decade, recently achieving human level performance on mathematics and programming benchmarks. In-spite of this, performance improvements on chemical tasks have emerged at a somewhat slower pace. In this work we investigate the capabilities of large language models (LLMs) in chemical search to address two central problems in AI-driven synthesis: retrosynthetic planning and mechanism elucidation. In our approach, the search environment builds options and the LLM serves as a guidance function to evaluate the validity and potential of a partially constructed solution. This is advantageous as LLMs can digest arbitrary inputs and optimize for arbitrary requirements. In this work, we show that LLMs can analyze and reason about chemical entities like molecules and reactions. We then leverage these capabilities in the context of two central problems in organic chemistry: retrosynthetic planning and mechanistic elucidation. Our results show that LLMs can accurately reason about chemical entities in both local and global terms, analyzing single reactions but also whole synthetic routes, and that such capabilities can be exploited through search algorithms for solving chemical problems in more flexible terms.

1 INTRODUCTION

Remarkable improvements in Large Language Models (LLMs) and their applications have been achieved in recent years across several domains. Applications in the natural sciences have emerged as an important research topic, as shown by the recent surge in benchmarks for scientific tasks (Phan et al., 2025; Mirza et al., 2024; Ruan et al., 2024; Guo et al., 2023), along with LLM-powered methods to tackle them (Dubey et al., 2024; Guo et al., 2025; OpenAI, 2023). In chemistry and materials science, applications span established challenges, like retrosynthetic planning (Guo et al., 2023), molecule and materials design (Grandi et al., 2025), and data extraction from the literature, as well as novel reformulations of existing problems, including literature question-answering (Q&A) (Lála et al., 2023), research lab assistants (Schmidgall et al., 2025), and autonomous systems (Boiko et al., 2023; M. Bran et al., 2024). Parallel to this, novel chemistry-specific benchmarks have emerged to assess LLMs' chemical knowledge through multiple choice questionnaires (Alampara et al., 2024).

Evidence demonstrates that current LLMs can display diverse types of chemical reasoning (Mirza et al., 2024). Furthermore, LLMs' reasoning patterns reflect more similarities to the way humans tackle problems in chemistry than traditional software (such as quantum mechanical calculations or other machine learning models); they can make reasonable assumptions, describe qualitative details of chemical entities, propose and develop ideas, among others (Guo et al., 2025). While ML models excel at predicting specific property values for novel compounds, human chemists —and similarly, LLMs— are better suited for analyzing reactions and synthetic strategies, proposing mechanisms, explaining modes of action, and reasoning about chemical trends, among others.

In that sense, it has been proposed that the greatest potential for LLMs in science lies in their ability to generate plausible hypotheses that can be tested, contrasted, applied and used to advance scientific understanding (Kumbhar et al., 2025; Cohrs et al., 2024; Zimmermann et al.). Some progress has been made recently in claim verification (Skarlinski et al., 2024; Trinh et al., 2024), contextual understanding of scientific literature (Lála et al., 2023), among others. However, a key limitation of these models is in the types of output they can produce. Sequences of text tokens can hardly

be translated into useful chemical objects such as molecules or mechanistic hypotheses. This is particularly evident in their poor performance at directly generating SMILES strings (Jang et al., 2024; Walters; Edwards et al., 2022; Christofidellis et al., 2023) making it impractical to use their raw outputs for tasks like synthetic planning or mechanism prediction in any meaningful way.

Building on these insights, we take a step back to find how hypotheses and explanations are typically ideated, articulated, and used to advance chemical understanding, along with the types of objects that are created for this goal. We propose a general methodology for decoding LLM's chemical knowledge into such objects, by leveraging search algorithms with language-driven heuristics. We demonstrate our approach in two critical application cases in organic chemistry, namely prompt-guided retrosynthetic planning, and mechanistic determination of organic reactions. Our results show that LLMs can effectively guide search processes and select optimal solutions from candidate lists based on query relevance. Furthermore, we provide insights into how both pretraining scaling (Kaplan et al., 2020) and test-time inference scaling (Snell et al., 2024; Guo et al., 2025) affect the quality and reliability of these results, establishing important practical considerations for deploying and improving such systems.

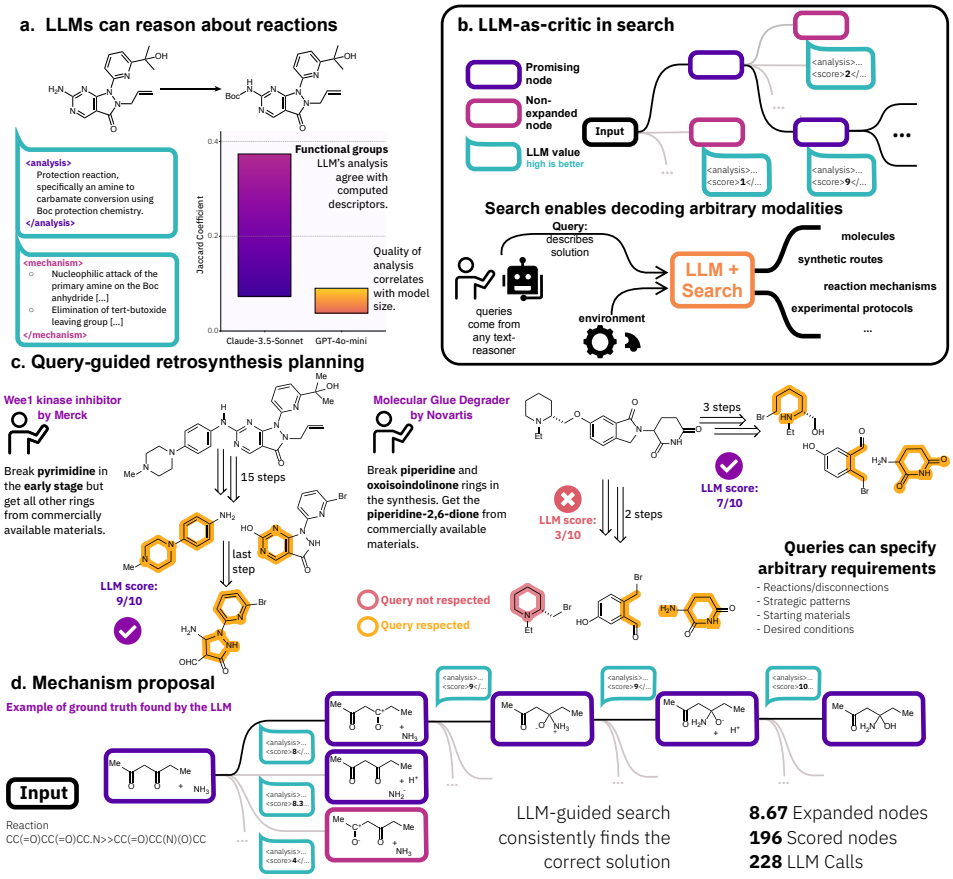

Figure 1: **a** LLMs are good at analyzing chemical entities like molecules and reactions. **b** These capabilities enable LLM-guided search, where models evaluate potential paths to prioritise node expansion. **c** Application to query-guided retrosynthesis planning aims to find a synthetic route for a target, and the query specifies describes properties of the desired route. LLM Scores reflect alignment of current solution with the query (0 to 10). **d** Application to mechanism determination, where LLMs guide the search for plausible mechanisms given reactants and products. The query can consist of specifications of conditions, results from experiments, or any other relevant context.

2 RELATED WORK

Search in chemistry Many important problems in chemistry are inherently search problems or can be formalised as such. Retrosynthetic planning is a canonical example, where the goal is to recursively disconnect a molecule into increasingly less complex precursors until commercially available materials are reached. Currently, the best methods construct libraries of known transformations —defining an action space— and then apply algorithms like MCTS (Browne et al., 2012; Segler & Waller, 2017) and A* (Hart et al., 1968; Chen et al., 2020), with policies and heuristic functions either carefully designed (Corey & Wipke, 1969; Corey et al., 1985; Grzybowski et al., 2023) or learned (Chen et al., 2020). While these systems have shown promising results (Genheden & Bjerrum, 2022; Torren-Peraire et al., 2024; Maziarz et al., 2023), they often struggle with novel reactions outside their training data and lack the chemical intuition that human experts use to prioritise promising synthetic routes (Schwaller et al., 2019; Fortunato et al., 2020). While most work in retrosynthetic planning focuses on the *general* case, potentially more impactful is the case of *goal-oriented* synthetic planning. This sets out to develop a synthesis pathway which algins to a *pre-defined goal*. For example, an expert chemist might choose to prioritise routes that leverage stocks of available starting materials (Armstrong et al., 2024; Yu et al., 2024), enforce late-stage ring construction, schedule the introduction of reactive groups only in the final steps, or ensure that hard-to-synthesise stereo centers and bonds are introduced via readily available precursors. Some of these principles are well established (Corey, 1967), however only recently techniques have been proposed to enable starting material and bond preservation or disconnection constraints (Thakkar et al., 2023; Westerlund et al., 2024; Yu et al., 2022; 2024; Armstrong et al., 2024). While displaying promising results, these methods still rely on the development of specalised systems for each new constraint, with an *arbitrary goal-oriented* synthesis tool remaining unexplored in the literature.

Another important problem suitable to search is the proposal of plausible reaction mechanisms. Here, the goal is to identify a sequence of elementary steps that explain the formation of products from reactants (Kayala & Baldi, 2012; Fooshee et al., 2018; Bradshaw et al., 2018). The importance of mechanisms is central to the understanding of chemical reactions (Cheng et al., 2015; Fey & Lynam, 2022), and typically multiple possible mechanisms can be conjectured for the same reaction (Stasiuk et al., 1956). Selection among these possibilities is traditionally conducted through experimental studies (Stasiuk et al., 1956) and supported by computational analyses (Yang et al., 2019; Glancy et al., 2020). Generating such mechanistic conjectures typically requires either brute-force search (Zimmerman, 2013; Zhao & Savoie, 2021), computationally-guided search (Bradshaw et al., 2018; Kayala & Baldi, 2012; Fooshee et al., 2018), or extensive human guidance (Herges, 1994). However these approaches struggle to scale for large systems or long reaction pathways.

LLMs in chemistry Applications of LLMs in chemistry range from traditional challenges like synthesis planning, reaction prediction, and condition recommendation, to property prediction and data extraction (Guo et al., 2023; Chen et al., 2023; Qian et al., 2023; Schilling-Wilhelmi et al., 2024). More novel applications include autonomous agents in robotic laboratories (Boiko et al., 2023; M. Bran et al., 2024) and research assistants (Darvish et al., 2024; Zheng et al., 2023). While these models demonstrate outstanding reasoning capabilities and chemical understanding, they face significant limitations in generating valid chemical representations, particularly SMILES strings, though they have shown promising performance in generating molecules and materials directly in cartesian space (Guo et al., 2023; Flam-Shepherd & Aspuru-Guzik, 2023).

LLMs and search Recent work has explored the synergistic combination of LLMs and search algorithms. (Schultz et al., 2024) distinguish between external and internal search with LLMs. Internal search refers to the LLM's own process of elaborating a reasoning path toward a solution, an approach exploited with techniques like chain-of-thought (Wei et al., 2022; Yao et al., 2022; Renze & Guven, 2024), and more recently baked into models with reinforcement learning techniques (Guo et al., 2025). External search, in contrast, integrates LLMs into traditional search algorithms, where they can serve as action generators(Ahn et al., 2022).

This integration has proven particularly effective across domains. In mathematics, LLMs have enhanced proof search by proposing auxiliary elements (Trinh et al., 2024). In chemistry, they have been successfully combined with Genetic Algorithms as mutation and crossover operators for molecular optimization (Wang et al., 2024), and integrated into Bayesian Optimization frameworks for op-

timizing both reactions and molecular properties (Ranković & Schwaller, 2023; Nguyen & Grover, 2024; Ye et al., 2025). These applications demonstrate how LLMs can provide sophisticated guidance while allowing traditional search algorithms to handle the structured exploration of solution spaces.

3 RESULTS

In this work we explore the synergy between search algorithms and LLMs to tackle two key problems in chemistry: goal-oriented retrosynthetic planning, and mechanistic determination for organic reactions. The first is a novel task where, in contrast to traditional works that focus on open-ended search of retrosynthetic paths, the input is a target molecule together with a natural language description of the desired disconnection approach, we refer to this task as prompt-guided retrosynthetic planning. This may include details about the desired types of transformations, assessments of feasibility, desired bond disconnections and synthetic stage at which reactions may occur, or more global descriptions of synthetic strategies, conditions, among others.

The second task is the generation of plausible mechanistic explanations of chemical reactions. This task is of crucial importance for chemists both in a practical and epistemological sense, as these are central for the mechanistic understanding of chemical reactions (Anslyn, 2006; Carey & Sundberg, 2007), which allows to infer parameters to optimise in chemical transformations. More fundamentally, mechanisms are used to derive hypotheses that might lead to the discovery of novel chemical transformations, a process that has resulted in several Nobel prize-winning discoveries (Woodward & Doering, 1945).

3.1 PROMPT-GUIDED RETROSYNTHETIC PLANNING

Recognizing LLM's capabilities for reaction analysis, we propose the use of LLMs as general scorers for guiding the search for synthetic routes with specific properties, as specified by an input query. For our experiments, we designed a benchmark where each task consists of pairs (SMILES, prompt), where the SMILES defines the synthetic target, and the prompt specifies desired properties of the solutions. Each of these tasks is accompanied by a rule-based scoring script, each specifically designed to rank synthetic routes according to the specific prompt; see Appendix A.1 for details.

The benchmark can be used in several ways: for reranking a set of pre-computed routes based on alignment with the query, or by directly performing search, then automatically scoring the resulting routes.

Route reranking For this variant of the problem, each task consists of a diverse set of precomputed routes, and the goal is to score each route based on their alignment with the task-specified query. As defined in the benchmark (see Appendix A.1.3), each task is accompanied by an evaluation script that automatically assigns an alignment score to each route. Performance of a given system is thus assessed as the correlation between these scores, and the scores generated by the evaluated system.

In this work, an instance of an LLM is prompted with the query along with a linearised version of the synthetic tree, which represents the full synthetic tree in a text format; see Appendix A.1.2. The LLM is prompted to analyse the inputs, and give a score between 0 and 10 that assesses the alignment of the given route with the query. The results in Figure 2.b show a clear advantage of Claude-3.5-Sonnet and DeepSeek-R1 over all the other tested LLMs, followed by DeepSeek-V3 and GPT-4o which also show good performances in some of the tasks. Smaller models, like GPT-4o-mini, show performances undistinguishable from random, indicating flaws in their chemical reasoning capabilities and knowledge, which links directly to their understanding of reaction SMILES and the query, as well as their capability to manifest and use chemical knowledge in real use-cases.

This task is already useful to assess model's knowledge and abilities to reason in chemistry, among other things. Despite of that, for a real application it still limited in that the fulfillment of the query relies on the assumption that a solution with the described properties already exists in a solution set, which is true of our benchmark but does not represent the general case. We thus now switch to the task of directly performing search that is biased towards the solutions with the particular properties described by a query.

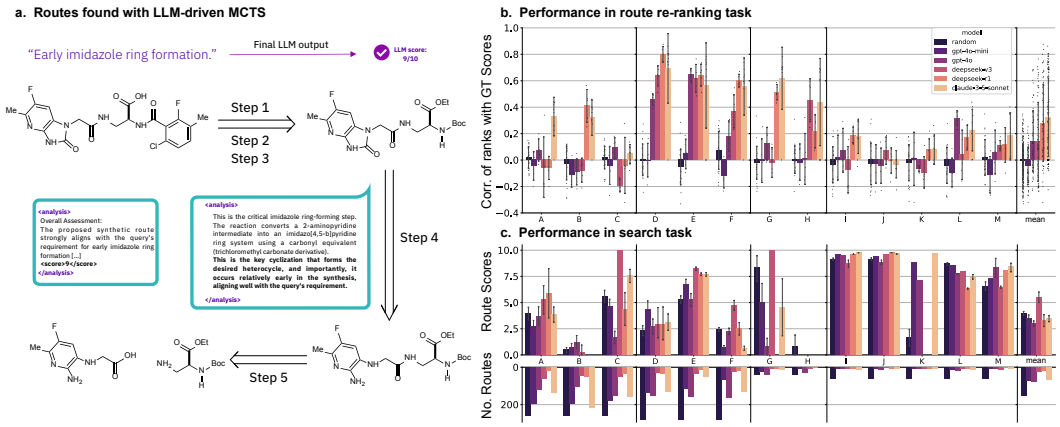

Figure 2: **a** Example of a synthetic route found by the proposed methodology. The method is evaluated on a benchmark designed as described in the Appendix A.1 in two tasks: **b** In route re-ranking, models are tasked with scoring the relevance of a pre-computed synthetic route, given a user query –a description of the desired solution. **c** MCTS Search, where models are used within an MCTS environment. At each step, models are queried to assess the value of potential nodes to expand, given a partially constructed route along with a user query.

LLM-powered search In this variant of the problem, we start from scratch with only a target and a query as defined by the task. The goal is then to directly generate synthetic routes that align closely with the query, while excluding other potential solutions that are less relevant. We implemented a modification to MCTS where an LLM is instantiated and serves as a complementary value function evaluating partially constructed routes and guiding the search toward solutions that align with the input query. At each node expansion step, the LLM assesses potential disconnections based on their likelihood of leading to a solution that satisfies the query constraints.

This guided exploration allows the search to focus on promising regions of the synthetic space that are more likely to yield routes matching the desired characteristics. As shown in Algorithm 3.1, the LLM-guided MCTS consists of the usual MCTS with UCB scores s calculated for each node, however at each expansion, a new value **H** –a heuristic value– is calculated, which consists of the LLM that analyses the current solution and each potential expansion with respect to the query q. The resulting scores $s + w * H$ are then used for selection, where w is a weight parameter used for balancing the magnitudes of s and H. For efficiency reasons, the LLM is only called randomly with probability p_{llm}.

Algorithm 3.1: LLM-guided MCTS

$lm \leftarrow LLM(q)$
while within budget **do**
  1. Select:
    $s \leftarrow UCB(node)$

  if $random() < p_{llm}$ **then**
    $H \leftarrow lm(children)$

  else
    $H \leftarrow \vec{0}$

  end if
    $child \leftarrow \arg\max(s + w \cdot H)$
  2. Expand: add new nodes
  3. Simulate: rollout
  4. Backpropagate: update values
end while

To evaluate the approach, we use the same tasks from benchmark A.1 as input. The resulting routes are then scored using each task's scoring script, and performance is measured as the average score of the resulting routes. A high value indicates that the resulting algorithm produces a set of routes that is rich in routes that align with the query. Results in Figure 2.c show the performance of several LLMs under this scenario. The results indicate that in general, the MCTS+LLM approach yields a low number of high-quality routes, as compared to the baseline which yields or-

ders of magnitude more routes, however not necessarily better routes. In contrast with the results in Figure 2.b, the performance gains against the baseline are rather limited. This is potentially due to a limited awareness of the scope or of longer-term planning of LLMs in the context of search. In particular, LLMs excel whenever the key element in the query happens early in the search, e.g. in "Late stage imidazole ring formation" (task E), the LLM identifies a step that aligns with the query (i.e. the ring formation) early in the search, and from there on success in search is well determined. In contrast, the tasks where the method fails the most are those where the query specifies requirements happening by the end of the search, e.g. "Early stage imidazole ring formation" (task F). Appropriately solving this requires specific planning skills, that allow the LLM to rate highly some partial routes although the key request doesn't *yet* happen.

3.2 Synthetic route analysis

While computational systems have made significant progress in *generating* synthetic routes, tools for *evaluating and analyzing* computationally generated synthesis pathways remain underdeveloped. Existing tools include similarity, route cost and route clustering (Genheden et al., 2021; Badowski et al., 2019; Genheden & Shields, 2025). However such measures typically rely on traditional *rule-based* chem-informatics or graph-theoretic functions.

Building on the promising results from Section 3.1, we extended our inquiry to explore additional features of synthetic routes that LLMs can interpret. In this effort, we compiled a set of complementary analytical use cases intended to highlight additional capabilities of LLMs in analysing entire synthesis pathways.

Starting material based semantic similarity In this analysis, we examine how Large Language Models (LLMs) comprehend synthetic strategy through semantic descriptions of starting materials used in chemical synthesis. Starting materials play a crucial role in determining synthetic strategy through providing pre-constructed structural motifs, functional groups and stereocenters to the synthesis pathway. By doing this, they effectively constrain the chemical space, determining which chemistry is feasible, and in what order. As such, an ability to parse, understand, and extract semantic value from starting materials can be viewed as a necessary pre-condition for LLM's to understand entire synthetic routes. By focusing on how LLMs process these basic chemical building blocks, we can better assess their capacity to comprehend more complex aspects of chemical synthesis.

We split this work into two studies based on comparing *rule-based* with the outputs from LLM's. Firstly, we ask the LLM to describe the synthetic route with relation to its starting materials. This description is embedded using OpenAI embedding models and a pairwise similarity matrix is constructed. We then extract the ground truth starting material SMILES from the relevant routes and construct an equivalent pairwise matrix using the size of the set intersection as a similarity measure. We find reasonable correlation between the two matrices, and display the plots with addition clustering on the embedding space in Figure A.6.

Seeking to obtain more detailed insight into LLM's understanding of starting materials we constructed a more powerful task. We construct an experimental setting where an LLM is tasked to extract all functional groups from the starting materials in a synthetic pathway. By doing so we directly assess whether LLMs grasp the chemical constraints that dictate the available reaction space and order of transformations. We use an in-house rule based system for functional group extraction to determine the ground truth set and measure (treating a functional group as a token) the LLM's error using the Jaccard co-efficient. Claude-3.5-sonnet shows higher overall alignment with the rule-based ground truth, while GPT-4o-mini generates more functional groups but suffers lower precision. Both LLM's show significant variance in their output. Minor formatting and naming differences between the rule-based and LLM outputs obscure direct comparisons, likely understating the true accuracy of the LLMs. We leave an anecdote that LLMs manage to correctly extract additional ring-system functional groups which are not currently tagged by our rule-based approach.

Case Study - Protecting Group Analysis Based on the previously established hypothesis that LLMs can effectively recognise and interpret the functional groups present in a synthetic route, we demonstrate this capability through qualitative, practical examples. A key characteristic of synthetic routes is the strategic use of protecting groups, which temporarily mask reactive functional groups to allow selective transformations. This enables chemists to perform selective transformations on

other parts of the molecule, with the protecting group removed afterward. This strategy, sometimes known as *tactical combinations*, allows the temporary complexity increase of protection to enable transformations that substantially advance the molecular structure. Correct understanding of protecting groups remains a major weakness of existing synthesis tools. These tools often propose either non-selective reactions that require protection, or conversely, include redundant protecting groups that add unnecessary steps.

We probe LLM understanding of this concept through a test using our benchmark set of synthetic routes. To assess the LLM's capabilities, we conduct two complementary experiments; firstly we investigate false positive detection by having LLMs identify routes where the base retrosynthesis tool has unnecessarily proposed protected groups and secondly, we investigate the inverse, flagging routes which require protecting groups but do not have them. Figures detailing this are given in A.2.1. In the false positive case, Claude-3.5-Sonnet correctly identifies an unnecessarily protected ethyl ester carboxylic acid in one of the starting materials which is then carried through and removed in the final step of the synthesis. The LLM demonstrates accurate reasoning in two key instances: First, it recognises that the initial amide bond formation does not require protection, given the substantially higher nucleophilicity of amines compared to carboxylic acids (Gromek et al., 2016). Second, it correctly determines that the penultimate step, involving phosgene-driven amide bond ring synthesis, does not require protection of free carboxylic acids due to both kinetic and entropic advantages inherent in five-membered ring formation (English et al., 1945; Clark & Pessolano, 1958). On the inverse task, the LLM tags a reactive hydroxyl group as potentially reacting intramolecularly in a polymerisation reaction with a bromide group elsewhere on the molecule, suggesting one of the two common hydroxyl protecting groups; TBS (tert-butyldimethylsilyl) or MOM (methoxymethyl).

Taken together, these case studies underscore how LLMs demonstrate a growing proficiency in understanding more complex aspects of chemical synthesis Their capacity to flag both unnecessary additions and omissions suggests promising avenues for refining the outputs of synthesis planning tools. This analytical momentum naturally leads us to the next challenge—mechanistic determination—where we harness similar LLM-guided strategies to explore and predict the elementary steps underlying chemical transformations.

3.3 MECHANISTIC DETERMINATION

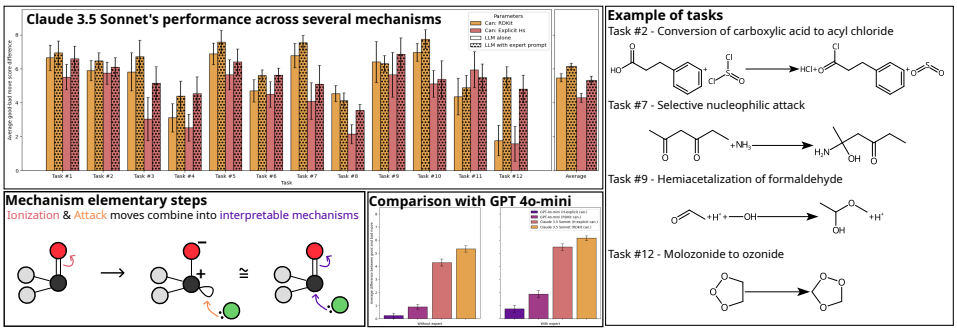

Figure 3: Results overview for mechanism search in a ionization/attack framework using LLMs as policy, and beam search as the search algorithm.

In chemistry, a reaction mechanism is a specification of why and how a given chemical change occurs, by means of a set of elementary steps (Clayden et al., 2012). The power of mechanisms in chemistry lies not only in its explanatory power of a single reaction instance, but also in that the reach of an explanation may extend further than only that reaction; potentially explaining more observed reactions, but also even predicting potential unknown transformations.

We follow a similar approach as for the task of synthetic planning - first evaluate the LLM's performance in a reranking or scoring task, and then switch to study its use in a search environment. We state the problem of finding a mechanism as a search problem with a limited set of predefined elementary steps, see Appendix A.3.1. A suitable search algorithm is then used to explore the space of possible mechanisms, while guided by the LLM to steer search towards solutions that make the most sense chemically.

Single-step scoring At search time, an LLM will be used to rank a list of possible next steps, given a partially constructed solution. The paths to explore and expand will be decided within a suitable search algorithm with the LLM serving to steer, as shown in Figure 1.

To assess the LLM's ability to select possible paths, a benchmark was designed that consists of a sample of N reactions, each with a mechanism approximated to use the single steps from Appendix A.3.1. Additionally, at each step in the mechanism (state), a set of 5 other intermediates reachable in 1 elementary step (but wrong ones) are generated. The task then consists of scoring each of the 6 possibilities, given the current state and each of the possible moves, and performance is measured as the difference between the LLM-given scores for the ground truth move and for the wrong moves, see Figure 3

This benchmark was used to ablate and compare multiple variations of LLM and prompting techniques, among others, and the results are shown in Figure 3

4 Conclusion

In this work we have demonstrated that large language models, when integrated with traditional search algorithms, offer a compelling new approach to tackling complex chemical problems such as prompt-guided retrosynthetic planning and mechanistic determination. By leveraging LLMs as evaluative agents within search frameworks—whether through re-ranking synthetic routes or guiding Monte Carlo Tree Search exploration—we have shown that these models can effectively translate abstract, language-based queries into actionable, chemistry-specific insights. In our studies, LLM-guided search consistently led to the identification of synthetic routes and reaction mechanisms that align closely with predefined query constraints, highlighting the models' capacity for chemical reasoning and strategic decision-making that mirrors expert intuition.

Overall, this work establishes a solid foundation for the integration of LLMs into chemical search tasks and underscores their potential to augment traditional cheminformatics methods. Future efforts will focus on further refining these techniques—through improved model scaling, task-specific fine-tuning, and enhanced prompt design—to unlock greater flexibility in synthetic planning.

In this work we demonstrate potential tasks in chemical synthesis which can be understood and partially addressed using large language models.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *arXiv preprint arXiv:2411.16955*, 2024.

EV Anslyn. Modern physical organic chemistry, 2006.

Daniel Armstrong, Zlatko Joncev, Jeff Guo, and Philippe Schwaller. Tango*: Constrained synthesis planning using chemically informed value functions. *arXiv preprint arXiv:2412.03424*, 2024.

Tomasz Badowski, Karol Molga, and Bartosz A Grzybowski. Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans. *Chemical science*, 10(17):4640–4651, 2019.

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

John Bradshaw, Matt J Kusner, Brooks Paige, Marwin HS Segler, and José Miguel Hernández-Lobato. A generative model for electron paths. *arXiv preprint arXiv:1805.10970*, 2018.

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

Francis A Carey and Richard J Sundberg. *Advanced organic chemistry: part A: structure and mechanisms*. Springer Science & Business Media, 2007.

Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. Retro*: Learning retrosynthetic planning with neural guided A* search. In *International Conference on Machine Learning*, pp. 1608–1616. PMLR, 2020.

K Chen, J Li, K Wang, Y Du, J Yu, J Lu, L Li, J Qiu, J Pan, Y Huang, et al. Chemist-x: Large language model-empowered agent for reaction condition recommendation in chemical synthesis, arxiv, 2023. *arXiv preprint arXiv:2311.10776*, 2023.

Gui-Juan Cheng, Xinhao Zhang, Lung Wa Chung, Liping Xu, and Yun-Dong Wu. Computational organic chemistry: bridging theory and experiment in establishing the mechanisms of chemical reactions. *Journal of the American Chemical Society*, 137(5):1706–1725, 2015.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. *Int. Conf. Mach. Learn.*, 2023.

Robert L Clark and Arsenio A Pessolano. Synthesis of some substituted benzimidazolones. *Journal of the American Chemical Society*, 80(7):1657–1662, 1958.

Jonathan Clayden, Nick Greeves, and Stuart Warren. *Organic chemistry*. Oxford University Press, USA, 2012.

Kai-Hendrik Cohrs, Emiliano Díaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. Large language models for causal hypothesis generation in science. *Machine Learning: Science and Technology*, 2024.

Elias James Corey. General methods for the construction of complex molecules. *Pure and Applied chemistry*, 14:19–38, 1967.

Elias James Corey and W Todd Wipke. Computer-assisted design of complex organic syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science*, 166(3902):178–192, 1969.

Elias James Corey, Alan K Long, and Steward D Rubenstein. Computer-assisted analysis in organic synthesis. *Science*, 228:408–418, 1985.

Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, et al. Organa: A robotic assistant for automated chemistry experimentation and characterization. *arXiv preprint arXiv:2401.06949*, 2024.

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.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *Proc. Conf. Empirical Methods Nat. Lang. Process.*, pp. 375–413, 2022.

JP English, RC Clapp, QP Cole, IF Halverstadt, JO Lampen, and RO Roblin Jr. Studies in chemotherapy. ix. ureylenebenzene and cyclohexane derivatives as biotin antagonists1. *Journal of the American Chemical Society*, 67(2):295–302, 1945.

Natalie Fey and Jason M Lynam. Computational mechanistic study in organometallic catalysis: Why prediction is still a challenge. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(4):e1590, 2022.

Daniel Flam-Shepherd and Alán Aspuru-Guzik. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files, 2023.

David Fooshee, Aaron Mood, Eugene Gutman, Mohammadamin Tavakoli, Gregor Urban, Frances Liu, Nancy Huynh, David Van Vranken, and Pierre Baldi. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering*, 3:442–452, 2018.

Michael E Fortunato, Connor W Coley, Brian C Barnes, and Klavs F Jensen. Data augmentation and pretraining for template-based retrosynthetic prediction in computer-aided synthesis planning. *J. Chem. Inf. Model.*, 60:3398–3407, 2020.

Samuel Genheden and Esben Bjerrum. Paroutes: a framework for benchmarking retrosynthesis route predictions. *ChemRxiv*, 2022.

Samuel Genheden and Jason D Shields. A simple similarity metric for comparing synthetic routes. *Digital Discovery*, 2025.

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. *J. Cheminf.*, 12:1–9, 2020.

Samuel Genheden, Ola Engkvist, and Esben Bjerrum. Clustering of synthetic routes using tree edit distance. *Journal of Chemical Information and Modeling*, 61(8):3899–3907, 2021.

John H Glancy, Daniel M Lee, Emily O Read, and Ian H Williams. Computational simulation of mechanism and isotope effects on acetal heterolysis as a model for glycoside hydrolysis. *Pure and Applied Chemistry*, 92(1):75–84, 2020.

Daniele Grandi, Yash Patawari Jain, Allin Groom, Brandon Cramer, and Christopher McComb. Evaluating large language models for material selection. *Journal of Computing and Information Science in Engineering*, 25(2):021004, 2025.

Samantha M Gromek, James A deMayo, Andrew T Maxwell, Ashley M West, Christopher M Pavlik, Ziyan Zhao, Jin Li, Andrew J Wiemer, Adam Zweifach, and Marcy J Balunas. Synthesis and biological evaluation of santacruzamate a analogues for anti-proliferative and immunomodulatory activity. *Bioorganic & medicinal chemistry*, 24(21):5183–5196, 2016.

Bartosz A. Grzybowski, Tomasz Badowski, Karol Molga, and Sara Szymkuć. Network search algorithms and scoring functions for advanced-level computerized synthesis planning. *WIREs Comput. Mol. Sci.*, 13(1):e1630, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

Rainer Herges. Coarctate transition states: The discovery of a reaction principle. *Journal of chemical information and computer sciences*, 34(1):91–102, 1994.

Hiroshi Hirai, Yoshikazu Iwasawa, Megumu Okada, Tsuyoshi Arai, Toshihide Nishibata, Makiko Kobayashi, Toshifumi Kimura, Naoki Kaneko, Junko Ohtani, Kazunori Yamanaka, Hiraku Itadani, Ikuko Takahashi-Suzuki, Kazuhiro Fukasawa, Hiroko Oki, Tadahiro Nambu, Jian Jiang, Takumi Sakai, Hiroharu Arakawa, Toshihiro Sakamoto, Takeshi Sagara, Takashi Yoshizumi, Shinji Mizuarai, and Hidehito Kotani. Small-molecule inhibition of wee1 kinase by mk-1775 selectively sensitizes p53-deficient tumor cells to dna-damaging agents. *Molecular Cancer Therapeutics*, 8(11):2992–3000, 2009.

Hyosoon Jang, Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. Can llms generate diverse molecules? towards alignment with structural diversity, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Matthew A Kayala and Pierre Baldi. ReactionPredictor: prediction of complex chemical reactions at the mechanistic level using machine learning. *J. Chem. Inf. Model.*, 52:2526–2540, October 2012.

John Ryan Kerrigan, Noel M. Thomsen, Artiom Cernijenko, Sarah E. Kochanek, Janetta Dewhurst, Gary O'Brien, Nathaniel F. Ware, Carina C. Sanchez, James R. Manning, Xiaolei Ma, Elizabeth Ornelas, Nikolas A. Savage, James R. Partridge, Andrew W. Patterson, Philip Lam, Natalie A. Dales, Simone Bonazzi, Sneha Borikar, Amelia E. Hinman, and Pamela Y. Ting. Discovery and optimization of first-in-class molecular glue degraders of the wiz transcription factor for fetal hemoglobin induction to treat sickle cell disease. *Journal of Medicinal Chemistry*, 67(22):TBD, 2024.

Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iquebal, and Chitta Baral. Hypothesis generation for materials discovery and design using goal-driven and constraint-guided llm agents. *arXiv preprint arXiv:2501.13299*, 2025.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

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*, pp. 1–11, 2024.

Krzysztof Maziarz, Austin Tripp, Guoqing Liu, Megan Stanley, Shufang Xie, Piotr Gaiński, Philipp Seidl, and Marwin Segler. Re-evaluating retrosynthesis algorithms with syntheseus. *arXiv preprint arXiv:2310.19796*, 2023.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, et al. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*, 2024.

Tung Nguyen and Aditya Grover. Lico: Large language models for in-context molecular optimization. *arXiv preprint arXiv:2406.18851*, 2024.

OpenAI. Gpt-4 technical report. *Preprint at https://arxiv.org/abs/2303.08774*, 2023.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Chen Qian, Huayi Tang, Zhirui Yang, Hong Liang, and Yong Liu. Can large language models empower molecular property prediction? *arXiv preprint arXiv:2307.07443*, 2023.

Bojana Ranković and Philippe Schwaller. Bochemian: Large language model embeddings for bayesian optimization of chemical reactions. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023.

Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.

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.

Mark Sabat, Daniel W. Carney, Gloria Hernandez-Torres, Tony S. Gibson, Deepika Balakrishna, Hua Zou, Rui Xu, Chien-Hung Chen, Ron de Jong, Douglas R. Dougan, Ling Qin, Simone V. Bigi-Botterill, Alison Chambers, Joanne Miura, Lucas K. Johnson, Jacques Ermolieff, Deidre Johns, Jangir Selimkhanov, Lily Kwok, Kevin DeMent, Chris Proffitt, Phong Vu, Erick A. Lindsey, Tony Ivetac, Andy Jennings, Haixia Wang, Padma Manam, Cipriano Santos, Cody Fullenwider, Rohan Manohar, and Andrew C. Flick. Design and discovery of a potent and selective inhibitor of integrin v1. *Journal of Medicinal Chemistry*, 67(12):TBD, June 13 2024.

Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T Koch, José A Márquez, and Kevin Maik Jablonka. From text to insight: large language models for materials science data extraction. *arXiv preprint arXiv:2407.16867*, 2024.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

John Schultz, Jakub Adamek, Matej Jusup, Marc Lanctot, Michael Kaisers, Sarah Perrin, Daniel Hennes, Jeremy Shar, Cannada Lewis, Anian Ruoss, Tom Zahavy, Petar Veličković, Laurel Prince, Satinder Singh, Eric Malmi, and Nenad Tomašev. Mastering board games by external and internal planning with language models, 2024.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.*, 5(9):1572–1583, 2019.

Marwin HS Segler and Mark P Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. Eur. J.*, 23:5966–5971, 2017.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Fred Stasiuk, WA Sheppard, and AN Bourns. An oxygen-18 study of acetal formation and hydrolysis. *Canadian Journal of Chemistry*, 34(2):123–127, 1956.

Amol Thakkar, Alain C Vaucher, Andrea Byekwaso, Philippe Schwaller, Alessandra Toniato, and Teodoro Laino. Unbiasing retrosynthesis language models with disconnection prompts. *ACS Central Science*, 9(7):1488–1498, 2023.

Paula Torren-Peraire, Alan Kai Hassen, Samuel Genheden, Jonas Verhoeven, Djork-Arné Clevert, Mike Preuss, and Igor V. Tetko. Models matter: the impact of single-step retrosynthesis on synthesis planning. *Digital Discovery*, 3:558–572, 2024.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Pat Walters. Silly Things Large Language Models Do With Molecules.

Haorui Wang, Marta Skreta, Cher-Tian Ser, Wenhao Gao, Lingkai Kong, Felix Strieth-Kalthoff, Chenru Duan, Yuchen Zhuang, Yue Yu, Yanqiao Zhu, et al. Efficient evolutionary search over chemical space with large language models. *arXiv preprint arXiv:2406.16976*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Annie M Westerlund, Lakshidaa Saigiridharan, and Samuel Genheden. Constrained synthesis planning with disconnection-aware transformer and multi-objective search. 2024.

Robert Burns Woodward and W von E Doering. The total synthesis of quinine. *Journal of the American Chemical Society*, 67(5):860–874, 1945.

Zhongyue Yang, Cooper S Jamieson, Xiao-Song Xue, Marc Garcia-Borràs, Tyler Benton, Xiaofei Dong, Fang Liu, and KN Houk. Mechanisms and dynamics of reactions involving entropic intermediates. *Trends in Chemistry*, 1(1):22–34, 2019.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *11th Int. Conf. Learn. Represent.*, 2022.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. Drugassist: A large language model for molecule optimization. *Briefings in Bioinformatics*, 26(1):bbae693, 2025.

Kevin Yu, Jihye Roh, Ziang Li, Wenhao Gao, Runzhong Wang, and Connor W Coley. Double-ended synthesis planning with goal-constrained bidirectional search. *arXiv preprint arXiv:2407.06334*, 2024.

Yemin Yu, Ying Wei, Kun Kuang, Zhengxing Huang, Huaxiu Yao, and Fei Wu. Grasp: Navigating retrosynthetic planning with goal-driven policy. *Advances in Neural Information Processing Systems*, 35:10257–10268, 2022.

Qiyuan Zhao and Brett M Savoie. Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. *Nature Computational Science*, 1(7):479–490, 2021.

Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062, 2023.

Xijun Zhu, Woong Sub Byun, Dominika Ewa Pieńkowska, Kha The Nguyen, Jan Gerhartz, Qixiang Geng, Tian Qiu, Jianing Zhong, Zixuan Jiang, Mengxiong Wang, Roman C. Sarott, Stephen M. Hinshaw, Tinghu Zhang, Laura D. Attardi, Radosław P. Nowak, and Nathanael S. Gray. Activating p53y220c with a mutant-specific small molecule. *bioRxiv*, 2024. doi: 10.1101/2024.10.23.619961. Preprint, not peer-reviewed.

Paul M Zimmerman. Automated discovery of chemically reasonable elementary reaction steps. *Journal of computational chemistry*, 34(16):1385–1392, 2013.

Yoel Zimmermann, Adib Bazgir, and Ben Blaiszik. Reflections from the 2024 large language model (llm) hackathon for applications in materials science and chemistry.

A APPENDIX

A.1 LANGUAGE-STEERED SYNTHESIS BENCHMARK

For our study's benchmark, we curated a dataset of four therapeutic molecules used in medicinal chemistry(Kerrigan et al., 2024; Sabat et al., 2024; Hirai et al., 2009; Zhu et al., 2024). We then augmented these molecules with strategically designed queries for the LLM. Unlike standard single-step templates often used in retrosynthetic planning, these queries combine multiple transformations into a single request and incorporate logic that can be cumbersome to encode directly.

More specifically, the queries address multiple ring-breaking transformations, avoid certain transformations by restricting starting materials to commercially available compounds, and consider the depth at which these transformations appear in the retrosynthetic search tree. It is important to note that our ring naming is not strictly aligned with IUPAC conventions; rather, it mirrors the way chemists informally discuss synthetic routes and highlight key substructures in natural language.

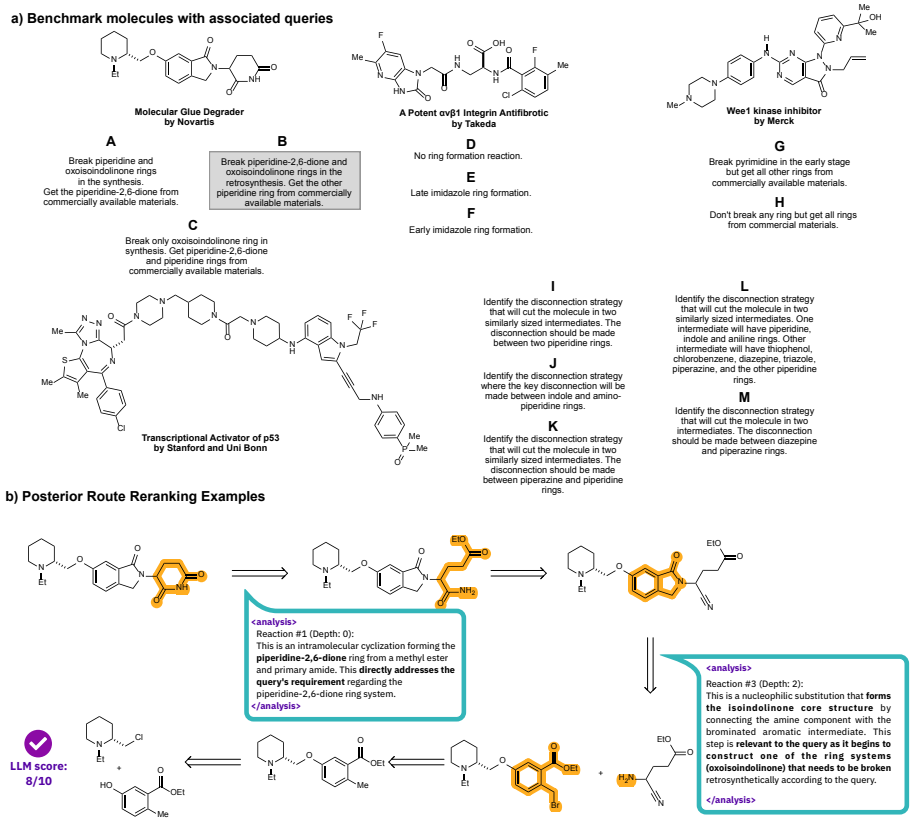

Figure A.1: a) Benchmark molecules and associated queries used for retrosynthetic reranking and search; b) Example of proposed synthetic route where LLM has correctly analysed two key query requirements.

A.1.1 PRECOMPUTED ROUTES FOR SYNTHESIS BENCHMARK

For the 4 targets shown in Appendix A.1, a set of routes was generated using an internally modified AiZynthFinder (Genheden et al., 2020) with internally trained policies and over 500 iterations, and maximum of 20 expansions per node. The resulting routes were then selected to adequately represent the multiple cases tested with the queries, so that each query is represented in the dataset with a representative percentage of the routes.

Type	SMILES	Value
RingBreakDepth	CC1=CC=C(Cl)C(C(N[C@H](C(O)=O)CNC(CN2C(C=C(F)C(C)=N3)=C3NC2=O)=O)=O)=C1F	58
RingBreakDepth	CN1CCN(C2=CC=C(N(C3=NC(N(C4=CC=CC(C(C)(C)O)=N4)N(CC=C)C5=O)=C5C=N3)[H])C=C2)CC1	39
MultiRxnCond	CCN1[C@@H](COC2=CC(C(N(C3C(NC(CC3)=O)=O)C4)=O)=C4C=C2)CCCC1	60
MultiRxnCond	CN1CCN(C2=CC=C(N(C3=NC(N(C4=CC=CC(C(C)(C)O)=N4)N(CC=C)C5=O)=C5C=N3)[H])C=C2)CC1	39
SpecificBondBreak	CP(C1=CC=C(NCC#CC2=CC(C(NC3CCN(CC(N4CCC(CN5CCN(C(C[C@@H]6N=C(C7=CC=C(Cl)C=C7)C(C(C)=C(C)S8)=C8N9C6=NN=C9C)=O)CC5)CC4)=O)CC3)=CC=C%10)=C%10N2CC(F)(F)F)C=C1)(C)=O	51-53

Table A.1: Molecular Data Summary

A.1.2 PROMPT FORMAT

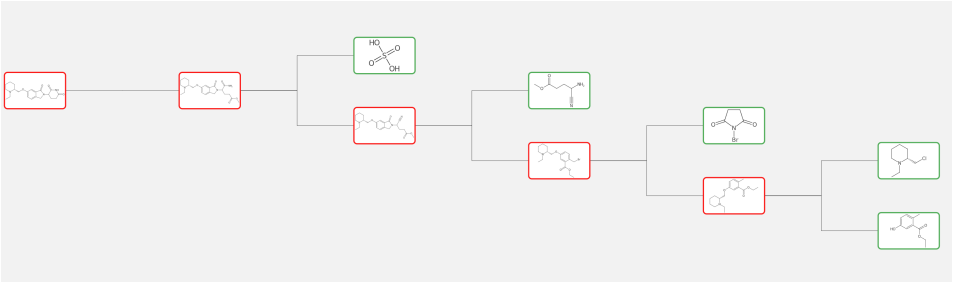

Figure A.2: a) Example of a synthetic route in Synthegy UI.

Retrosynthetic Analysis Query

You are an experienced organic chemist tasked with assessing the relevance or similarity of a proposed synthetic route to a given query. You will analyze the reactions carefully, explain the key points of each reaction in relation to the query, and then assess the relevance of the proposed plan for the given query.
The query provides a desired synthetic pathway towards a target molecule:

```
<query>
Break piperidine-2,6-dione and oxoisoindolinone rings [...]
</query>
```

Next, you will be given a sequence of proposed reactions, starting from the target molecule and going backwards through each of the intermediate reactions in a retrosynthetic way.
Note:

- ”Early” in the synthesis means further from the target molecule, as the reactions further back in the sequence are closer to the starting materials.
- ”Late” and ”late-stage” means closer to the target molecule.
- ”Break” indicates a retrosynthetic step, where a molecule is broken down into simpler components. In the forward direction, this would mean ”Form”. For example, ”Break C-C bond” would be equivalent to ”Form C-C bond” in the forward direction.

Each reaction is numbered and has a depth value indicating its position in the retrosynthetic tree:
Analyze each reaction in the proposed sequence, starting from the last one (closest to the product) and moving backwards. For each reaction:

- Identify the key functional groups and structural changes involved.
- Evaluate how well the reaction aligns with the query’s requirements.

Write your analysis for each reaction in separate `<analysis>` tags. Be sure to reference specific aspects of the query when discussing relevance.

```
Reaction #1. Depth: 0
[CH3:1][CH2:2][N:3]1[CH2:4][CH2:5][CH2:6][CH2:7]...
.
.
Reaction #5. Depth: 4
[CH3:1][CH2:2][N:3]1[CH2:4][CH2:5][CH2:6][CH2:7]...
```

After analyzing all reactions, assess the overall relevance of the proposed synthetic route to the query. Consider:

- How well does the overall sequence align with the query’s goals?
- Are there any major discrepancies or missing steps?

Provide a detailed justification for your assessment, drawing on your analysis of individual reactions and your expertise as an organic chemist.
Finally, assign a relevance score from 0 to 10, where 10 indicates the highest relevance to the query. Present your score in the following format:

```
<score>[integer from 0 to 10]</score>
```

Final Notes:

- The reactions shown are theoretical and have not been tested in a laboratory. They represent desired transformations but may not necessarily reflect what would actually occur in a flask.
- Your expertise is crucial in assessing the relevance of these proposed reactions.

A.1.3 RULE-BASED ROUTE SCORING

Given a synthetic route and a query, we need to provide a score that indicates how aligned the route is with the requirements in the query. For computing this, we classified queries into 3 types: single ring-breaking, single bond breaking specification, and multi-reaction specification.

The scoring then happens generally by traversing the synthetic tree and checking whether any reactions match a specified pattern depending on the type of query. As all the queries in our benchmark have a temporal aspect to them (early, late, key step, see Appendix A.1), the position at which the query matches, if any, is also recorded, and a score is calculated with these data. Thus every query requires the definition of 2 things: the matching condition, and the scoring function. Here we show the definitions for all 3 types of queries:

Snippet 1 Implementation of the RingBreakDepth scoring class for evaluating ring-breaking reactions at specific depths in synthetic routes.

```python
class RingBreakDepth(BaseScoring):
    def route_scoring(self, x) -> float:
        """x: depth at which condition is met in route / length of route."""
        if self.condition_type == "bool":
            if self.target_depth == -1:  # Positive if condition not met
                return 1 if x < 0 else 0
        else:
            if x < 0:
                return 0
            return abs(x - self.target_depth)

    def hit_condition(self, d):
        """We're looking specifically for ringbreaking(forming) reactions."""
        return d.get("metadata", {}).get("policy_name") == "ringbreaker"
```

Snippet 2 Implementation of the SpecificBondBreak scoring class for evaluating bond breaking specifications in synthetic routes.

```python
class SpecificBondBreak(BaseScoring):
    def __init__(self, config):
        """Bond to break is specified in benchmark file"""
        self.atom_1 = config["bond_to_break"]["atom_1"]
        self.atom_2 = config["bond_to_break"]["atom_2"]

    def route_scoring(self, x):
        """Disconnection happens (!=-1), + should happen late-stage roughly."""
        if x < 0:
            return 0  # Worst case - disconnection doesn't happen
        else:
            return (1 - x)  # Disconnection happens late-stage. The smaller x, the better.

    def hit_condition(self, d):
        """Determine if the bond between A1 and A2 is broken in current reaction."""
        rxn = d["metadata"]["mapped_reaction_smiles"].split(">>")
        prod = Chem.MolFromSmiles(rxn[0])
        reacts = [Chem.MolFromSmiles(r) for r in rxn[1].split(".")]

        if (self.atom_1 in [a.GetAtomMapNum() for a in prod.GetAtoms()]) and \
           (self.atom_2 in [a.GetAtomMapNum() for a in prod.GetAtoms()]):
            for r in reacts:
                if (self.atom_1 in [a.GetAtomMapNum() for a in r.GetAtoms()]) ^ \
                   (self.atom_2 in [a.GetAtomMapNum() for a in r.GetAtoms()]):
                    return True
        return False
```

Snippet 3 Implementation of the MultiRxnCond class for evaluating multiple reaction conditions involving various heterocyclic structures. The class MultiRxnCondBase is simply an extension of the BaseScoring class, to detect multiple conditions across different children in a synthetic tree.

```python
class MultiRxnCond(MultiRxnCondBase):
    def __init__(self, config):
        self.allow_piridine = config.get("allow_piridine") or False
        self.allow_piperazine = config.get("allow_piperazine") or False
        ...

    def condition_depth(self, d) -> Tuple[bool, int]:
        """Extract all the reactions from tree, and find if condition is met."""
        piridine = any(
            self.detect_specific_break(r, "c1ccncc1") for r in reactions
        )
        piperazine = any(
            self.detect_specific_break(r, "C1CNCCN1") for r in reactions
        )

        condition = (
            piridine == self.allow_piridine
            and piperazine == self.allow_piperazine
            and ...
        )
        return condition, len(reactions)
```

A.2 LLM ANALYTICAL CAPABILITIES IN ORGANIC SYNTHESIS ROUTES

A.2.1 PROTECTING GROUP ANALYSIS

Here we show the routes discussed in 3.2.

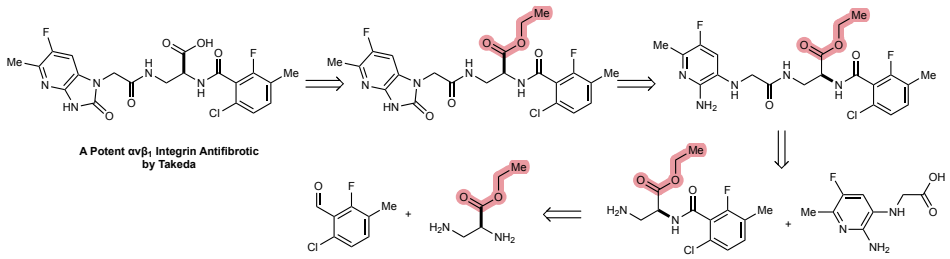

Figure A.3: Example route where the LLM correctly reasons that the highlighted ethyl ester protecting groups is superfluous to the route. We highlight the following papers for examples of similar chemistry working without carboxylic acid protection (English et al., 1945; Clark & Pessolano, 1958; Gromek et al., 2016)

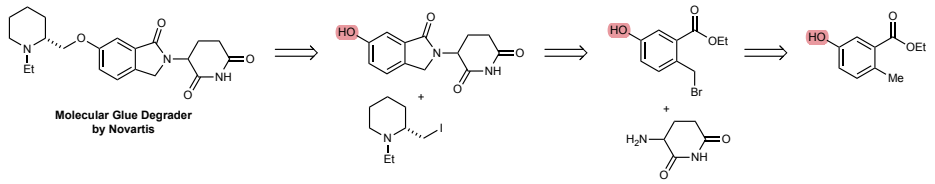

Figure A.4: Example route where the LLM correctly identifies the need for a protecting group on the highlighed hydroxyl functional group. Being a nucleophile the hydroxyl group could cause a competing intermolecular polymerisation reaction with the bromide group in the second step. Claude-3.5-sonnet generated description : "The phenol group should be protected before the bromination step to prevent side reactions and ensure selective bromination at the benzylic position. A suitable protecting group such as TBS (tert-butyldimethylsilyl) or MOM (methoxymethyl) would be recommended to mask the phenol during the bromination step and could be removed after the transformation is complete."

A.2.2 FUNCTIONAL GROUP EXTRACTION

Aiming to evaluate the understanding modern LLM's have about functional groups within *specific* molecules (starting materials) of a synthetic route we first use an in-house *rule-based* tool to extract a ground truth set of FG's from the starting materials of each route. This tool uses a list of 188 functional group SMARTS and performs substructure matching via RDKit to parse the functional from a given molecule. We then pass a linearised synthetic pathway to the LLM and ask it to provide a list of functional groups present in the starting materials. To ensure proper formatting, we include the names of the 188 starting materials in the prompt. We post-process the outputs by setting to lower case, sorting the words in a functional group phrase (chloride acyl is converted to acyl chloride) and measure Jaccard overlap between the LLM generated FG's. The Maximum, Minium and 25th to 75th percentiles are plotted in A.2.2. Claude-3.5-sonnet performs better on average. Below we provide example descriptions for the first route.

- Rule based : alkyl cl halide prim, alcohol tertiary, amine primary, any aryl bromide, ether, amine secondary, amine tertiary, silylchloride.
- Claude-3.5-sonnet : amine primary, amine secondary, aryl chloride, aryl bromide, alkyl halide, alcohol, ether, ester, chloride phosphoryl, chloride silyl, heterocycle, amide, lactam.

- GPT-4o-mini : alkyne terminal, amine secondary, amine tertiary, amide secondary, acyl halide, enolizable ketone, chlorine, phosphorylchloride px4, silicon, alcohol, ester nonenolizable, aldehyde enolizable, nitrile nonenolizable, thiophenol, nh sulfonamide, alcohol allyl terminal, alcohol allyl any, acyclic amide weinreb, amide cyclic weinreb, ewg ketimine, isocyanate, thioether, thioamide, aldehyde general thioacetal, general ketone thioacetal, hemiacetal, amide ether hemiaminal, thiourea, sulfonylchloride, and more.

We note that while gpt-4o-mini actually yields substantially higher *recall* at the expense of *precision*, as it tends to generate a far larger number of functional groups than those actually present in the starting materials. While formatting differences between the rule-based and LLM systems persist, we ignore them for this analysis as it likely *understates* the knowledge LLM's have about functional groups.

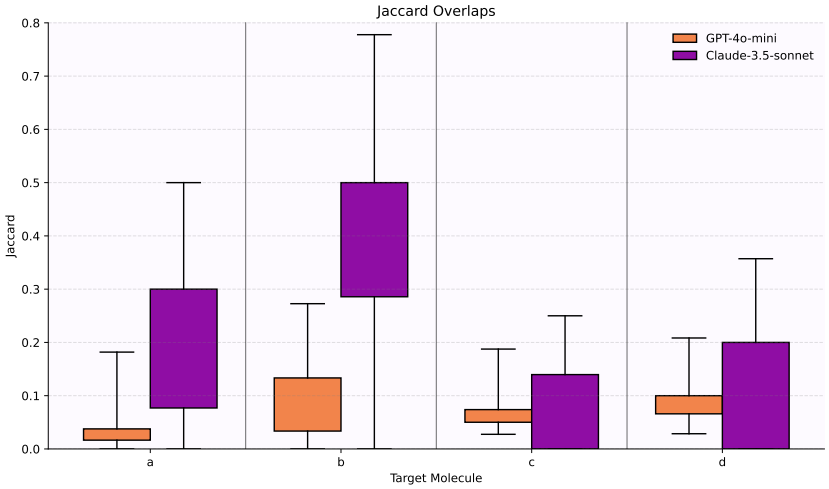

Figure A.5: Jaccard overlap between LLM and rule based extraction of functional groups from the starting materials of a synthetic route.

A.2.3 Starting material latent space clustering

Here we show pairwise similarity matrices comparing LLM description latent spaces and starting material set overlap. The LLM latent spaces were generated by having Claude-3.5-sonnet describe the starting materials in each route, with descriptions embedded using text-embedding-3-small. Starting material set overlap was computed by extracting all leaf nodes from each synthesis pathway and converting them into sets. Each row-column pair in the matrix represents a pairwise comparison using the same metric. Correlation between LLM descriptions and set overlap was calculated at the matrix level.

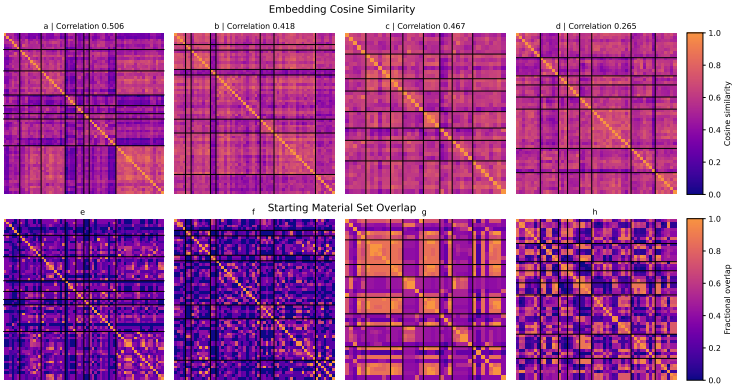

Figure A.6: Similarity matrices for the embeddings of LLM generated descriptions of starting materials and ground truth starting material set overlap.

A.3 MECHANISM PROPOSAL

A.3.1 ELEMENTARY MECHANISTIC STEPS

For the sake of completeness, but also of tractability, the mechanisms have been broken down into their elementary components. The possible actions at each state (corresponding to a set of molecules) consist of two fundamental types: ionization moves, in which any bond of the set of molecules decrease in bond order by one and ionise on any of its terminal atoms, and attack moves, in which any atom with a lone pair can attack any atom with an empty orbital, therefore increasing the bond order by one. Even though this set of moves is minimalistic, it has proven to be quite practical as a systematic way to enumerate and also to translate the majority of non-radical chemistry. However, a limit has been found regarding concerted moves. Even though such cases can often be inferred from the order and places in which following transitions happen, cases like SN1 vs. SN2 explicitly require one to know if the electron moves are concerted or not.

A.3.2 BENCHMARK

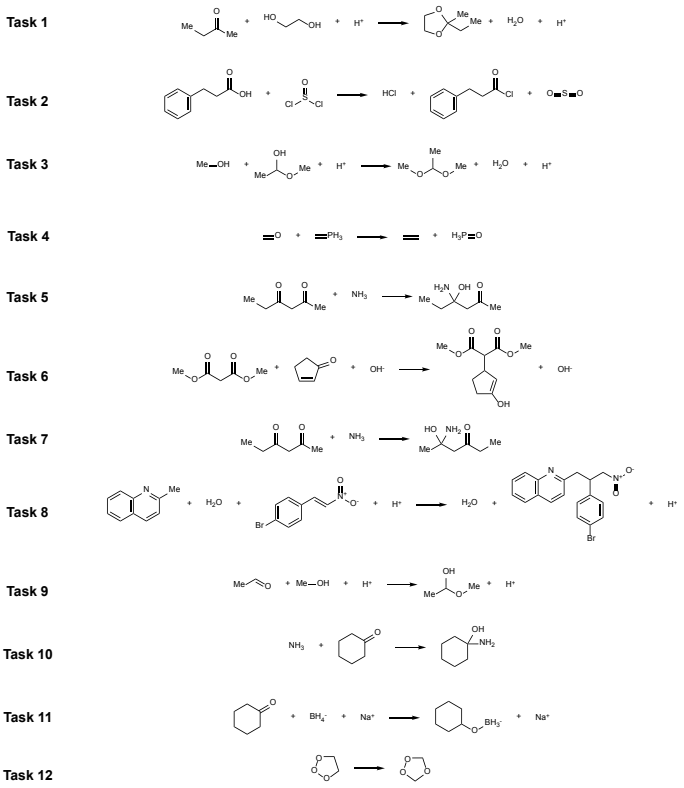

Figure A.7: All tasks given to for mechanism elucidation