ATOM-ANCHORED LLMS SPEAK CHEMISTRY: A RETROSYNTHESIS DEMONSTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Applications of machine learning in chemistry are often limited by the scarcity and expense of labeled data, restricting traditional supervised methods. In this work, we introduce a framework for molecular reasoning using general-purpose Large Language Models (LLMs) that operates without requiring labeled training data. Our method anchors chain-of-thought reasoning to the molecular structure by using unique atomic identifiers. First, the LLM performs a one-shot task to identify relevant fragments and their associated chemical labels or transformation classes. In an optional second step, this position-aware information is used in a few-shot task with provided class examples to predict the chemical transformation. We apply our framework to single-step retrosynthesis, a task where LLMs have previously underperformed. Across academic benchmarks and expert-validated drug discovery molecules, our work enables LLMs to achieve high success rates in identifying chemically plausible reaction sites ($\geq 90\%$), named reaction classes ($\geq 40\%$), and final reactants ($\geq 74\%$). Beyond solving complex chemical tasks, our work also provides a method to generate theoretically grounded synthetic datasets by mapping chemical knowledge onto the molecular structure and thereby addressing data scarcity.

1 INTRODUCTION

General-purpose large language models (LLMs) have advanced rapidly in recent years, finding increasing application in the domain of chemistry. A prominent example of this trend is the use of LLMs like GPT-4 Achiam et al. (2023) as high-level reasoning agents that leverage specialized chemistry tools to automate complex tasks Boiko et al. (2023); M. Bran et al. (2024). In this paradigm, the LLM orchestrates tool calls that encapsulate chemical logic and subsequently reasons over the tool outputs.

Beyond the use of general-purpose models, prevailing approaches either train specialized chemistry LLMs or adapt general-purpose LLMs to the chemical domain, where molecular data is represented in the Simplified Molecular Input Line Entry System (SMILES) format Weininger (1988); Weininger et al. (1989), a chemical notation for representing chemical graph structures as computer-readable strings. Examples of specialized chemistry LLMs include models that are solely pre-trained on SMILES data and then either fine-tuned for a specific downstream task (e.g., Ross et al. (2022); Irwin et al. (2022)) or used to extract molecular embeddings for downstream tasks (e.g., Ross et al. (2022); Sadeghi et al. (2024); Masood et al. (2025)). Alternatively, general-purpose LLMs are adapted to the chemical domain through methods such as supervised fine-tuning (SFT) Kim et al. (2024); Cavanagh et al. (2024), preference optimization (PO) Cavanagh et al. (2024), or the direct extraction of task-specific embeddings from general-purpose LLMs Sadeghi et al. (2024). Finally, recent work adapts Chain-of-Thought (CoT) Wei et al. (2023) chemistry reasoning models following the Deepseek-R1 Guo et al. (2025) paradigm, e.g., ether0 Narayanan et al. (2025) fine-tunes Mistral-Small-24B-Instruct Mis using SFT on Deepseek-R1 reasoning traces and PO on chemistry tasks.

However, a central challenge in chemical machine learning is the scarcity and high cost of labeled data. This presents a significant limitation, as the aforementioned approaches all rely on labeled data for model training. Nevertheless, recent studies have shown that general-purpose LLMs are capable of reasoning over chemical structures, yet this capability is often exercised indirectly. For instance,

general-purpose LLMs have been used to enrich SMILES with text descriptions to fine-tune smaller models Qian et al. (2023), address diverse chemistry tasks via zero-shot and few-shot prompting with varying success Guo et al. (2023), and solve chemical mathematical calculations by generating and refining code-based solutions Ouyang et al. (2024). A final category of applications addresses synthesis planning, the task of identifying viable synthetic routes by deconstructing a target molecule into smaller precursors using reactions until a set of commercially available starting materials is found Segler et al. (2018); Corey & Cheng (1989). In this context, LLMs can reason about chemical structures to guide and evaluate the synthesis planning process itself based on a desired provided route outcome prompt, without directly manipulating the structures Bran et al. (2025). As LLMs tend to struggle with generating high-quality reaction predictions directly, they can be paired with an evolutionary algorithm to reason over and evolve a population of full synthesis routes Wang et al. (2025). To ensure chemical validity, this process uses a database of known reactions and molecule routes, which are queried via a nearest-neighbor search in an embedding space to identify structurally similar precedents for chemical grounding.

In this work, we build on these insights to introduce a framework that enables general-purpose LLMs to successfully reason directly over molecular structures. Our method works by anchoring the reasoning process to a molecule's atom-maps, which are unique identifiers for each atom in a molecular SMILES. This approach mirrors a chemist's workflow, **operates without labeled training data or task-specific model training**, and consists of two stages. First, in a zero-shot task, the model performs a chemical analysis on the chemical structure to identify the atom-maps of relevant fragments for the task and assigns structural labels for these fragments solely based on chemical reasoning. Second, in an optional few-shot task, it transforms the chemical structure based on these identified fragments, guided by examples from a specific chemical transformation class (e.g., a particular reaction or other defined chemical transformation).

We apply this framework to single-step retrosynthesis, where the goal is to identify, given a product molecule, a set of plausible reactant molecules (precursors) that can form the product in a single reaction step Torren-Peraire et al. (2024). Formally, the goal is to learn a function $f(P) \rightarrow [R_1, R_2, \ldots, R_n]$ that maps a product molecule P to a ranked list of plausible reactant sets, $[R_1, R_2, \ldots, R_n]$, where each R_i is a set of one or more reactant molecules, $\{r_1, r_2, \ldots\}$, proposed to synthesize P. In this task, prior research shows that general-purpose LLMs are not competitive with specialized models as they underperform their specialized counterparts by more than 40 percentage points in top-1 accuracy Guo et al. (2023) or solve only one out of five test examples correctly Li et al. (2025). Our approach marks a shift from conventional supervised methods, which either (1) directly map products to reactants using Transformers Irwin et al. (2022); Tetko et al. (2020), Graph Neural Networks Chen & Jung (2021); Zhong et al. (2023), Markov Bridges Igashov et al. (2024), or fine-tuned LLMs Yang et al. (2024); Nguyen-Van et al. (2024), or (2) use a two-step, disconnection-aware paradigm where a model first learns to identify a bond disconnection site and second applies a transformation afterward. Our approach evolves the second paradigm. Whereas these supervised methods apply a learned mapping by selecting a site either automatically Thakkar et al. (2023); Kreutter & Reymond (2023) or with human guidance Thakkar et al. (2023); Westerlund et al. (2025), our work introduces explicit chemical reasoning as the core mechanism for both steps, leading to the following key contributions:

1. We introduce a novel reasoning framework that enables LLMs to zero-shot analyze and few-shot transform molecular structures without task-specific training by anchoring their reasoning process directly to the molecule's SMILES atom maps, thereby eliminating the need for labeled training data or task-specific model training.

2. We demonstrate the framework's effectiveness in single-step retrosynthesis on both academic benchmarks and expert-validated real drug discovery molecules, where it successfully identifies strategic disconnections, executes the corresponding transformation to predict reactant structures, and provides a chemically-grounded, explainable rationale for its predictions.

3. We establish a general blueprint for applying LLMs to data-scarce problems in computational chemistry, enabling the generation of novel synthetic datasets by mapping general chemical knowledge directly onto molecular structures.

2 METHODS

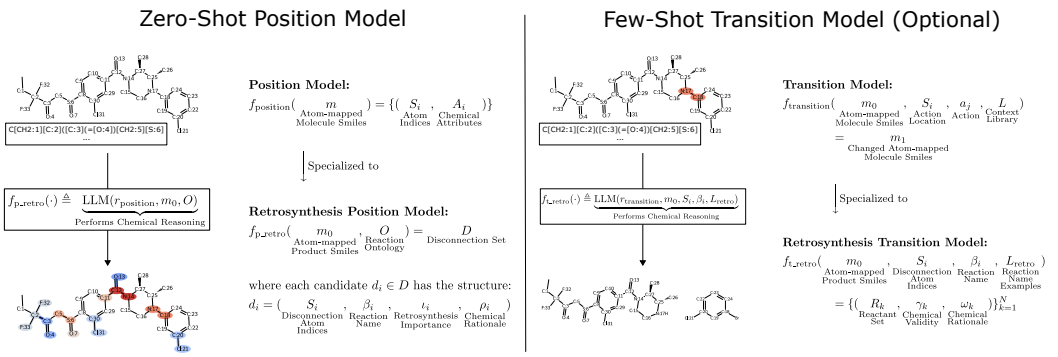


Figure 1: Adaptation of our general framework to the task of retrosynthesis. First, the Zero-Shot Position Model ($f_{\text{position\_retro}}$ or $f_{\text{p\_retro}}$, guided by r_{position}) analyzes an atom-mapped product m_0 together with the reaction ontology O to identify and rank disconnection candidates $(S_i, \beta_i, \iota_i, \rho_i)$. Second, the (optional) Few-Shot Transition Model ($f_{\text{transition\_retro}}$ or $f_{\text{p\_retro}}$, guided by $r_{\text{transition}}$ and a library L_{retro} of β_i reaction examples) applies the selected reaction β_i at the site S_i to generate plausible reactant molecules (R_k) with validity assessment (γ_k) and chemical rationale (ω_k).

2.1 FRAMEWORK

Conventional drug discovery models learn a direct mapping $f : \mathcal{X} \to \mathcal{Y}$, treating molecular representations $x \in \mathcal{X}$ as abstract data points to predict properties $y \in \mathcal{Y}$. This paradigm disregards the underlying chemical knowledge that could govern the relationship r between a molecule's structure and its properties. In contrast, our approach circumvents this data-driven mapping by leveraging the emergent reasoning capabilities of a pre-trained LLM. Guided by a natural language prompt, the LLM performs a detailed chemical analysis with its reasoning explicitly anchored to the molecule's SMILES atom maps, ensuring a precise linkage to specific structural locations. This structurally-grounded analysis enables the direct inference of chemical properties, eliminating the need for task-specific fine-tuning. Our approach operates in two stages:

1. **Zero-Shot Structural Analysis and Property Prediction (Position Model):** Guided by a natural language prompt $r_{position}$ that encodes domain knowledge about the task, the LLM analyzes an atom-mapped molecule SMILES m to identify relevant substructures. Based on this prompt-guided reasoning, which is explicitly linked to atom map indices, the position model $f_{position}(m)$ predicts a set of properties $P = \{p_1, \ldots, p_n\}$. Each prediction p_i is a tuple $p_i = (S_i, A_i)$, where $S_i \subseteq V(m)$ is a set of atom indices from the molecule m (the structural label), and $A_i = (a_1, a_2, ..., a_k)$ is an ordered tuple of inferred chemical attributes relevant to the task (e.g., "toxic," "reaction"). Each individual attribute a_i in this tuple can be a passive descriptor or an actionable transformation.
2. **Prompt-Guided Molecular Transformation (Transition Model):** In an optional second phase, predictions $p_i = (S_i, A_i)$ containing an actionable transformation in their attribute tuple A_i are executed. For each general chemical task, a transformation function $f_{transition}$ is defined by a second natural language prompt $r_{transition}$. This transition function executes an actionable attribute $a_j \in A_i$ by applying $f_{transition}$ to an initial molecule m_0 at the location S_i to yield a new molecule m_1, such that $m_1 = f_{transition}(m_0, S_i, a_j, L)$. Here, L is a context library providing examples or any relevant information for the established chemical operations identified by the actionable attribute a_i from the tuple A_i. This is feasible because many chemical transformations are discrete, well-established operations, allowing in-context learning to ensure chemical validity.

2.2 A POSITION MODEL FOR RETROSYNTHESIS

The Position Model emulates a human chemist's analytical workflow to identify and rank potential disconnection sites in a product molecule. Formally, given an atom-mapped product molecule m, the

Position Model is a function $f_{position\_retro}(m)$ that predicts a set of potential retrosynthetic disconnection candidates, $D = \{d_1, d_2, \ldots, d_N\}$. Each candidate $d_i = (S_i, \beta_i, \iota_i, \rho_i)$, which instantiates the general property prediction $p_i = (S_i, A_i)$ for retrosynthesis, is generated by the function:

$$D = \{(S_i, \beta_i, \iota_i, \rho_i)\}_{i=1}^{N} = f_{position\_retro}(m_0, O)$$

This function maps a set of inputs:

- m_0: The atom-mapped target product molecule canonicalized SMILES.
- O: A reaction ontology containing reaction names corresponding to a library of executable transformations L, providing a bridge to the optional transformation phase.

to a set of N distinct tuples:

- $S_i \subseteq V(m)$ is the structural label: a set of atom indices defining the disconnection point.
- β_i is the predicted reaction name: a chemical attribute identifying a suitable transformation (e.g., "Suzuki Coupling"). To make this actionable, we ground predictions using the reaction ontology (O), but do not strictly constrain them, allowing the suggestion of reactions outside of O (which are flagged).
- $\iota_i \in \mathbb{R}$ is the retrosynthesis importance: a score ranking the strategic value of the disconnection, which can be used to prioritize the most promising reactions (e.g., major ring-forming reactions, core scaffold construction).
- ρ_i is the chemical rationale: a text-based justification tied to primary strategic goals of retrosynthesis (e.g., structural simplification, reaction robustness, and stereochemical control).

The entire reasoning process of $f_{position\_retro}$ is defined by a natural language prompt $r_{position}$ (see Prompt 1). Crucially, $r_{position}$ does not contain explicit transformation rules (e.g., SMARTS patterns) or any other reaction-specific rules. Instead, it instructs the LLM to emulate a chemist's analytical workflow. Reframing the retrosynthesis task necessitates a shift in evaluation, moving beyond classical top-n performance based on product-reactant replication. Our evaluation instead measures the model's ability to correctly identify the ground-truth disconnection site and reaction type, for which the following metrics are defined:

1. Partial Match Accuracy: An indicator metric that is true if any predicted disconnection $S_i \in D$ has a non-empty intersection with the ground truth S_{gt}.
2. Best Match Jaccard: The highest Jaccard similarity between any predicted structural label $S_i \in D$ and the ground truth set S_{gt}.
3. Exact Match Accuracy: A stricter metric that is true if the best-matching predicted disconnection site (by Jaccard score) is identical to the ground truth S_{gt}.
4. Conditional Reaction Accuracy: Conditional on a partial match and the highest Jaccard similarity in D, this metric evaluates the reaction name(s) β_i from the disconnection candidate(s) d_i. The metric is 1 if any of these β_i match the ground truth reaction name, β_{gt}.

2.3 A Transition Model for Retrosynthesis

To complete the retrosynthesis workflow, we define the Transition Model as $f_{\text{transition\_retro}}$. This model uses a disconnection candidate d_i and a target product m_0 to generate a set of plausible reactants R. To simulate a chemist's literature lookup for a reaction, the reaction name $\beta_i \in O$ is used to sample up to five reaction examples from a training dataset to create the task-specific, in-context library L_{retro}. The one-to-many Transition Model is then defined as:

$$\{(R_k, \gamma_k, \omega_k)\}_{k=1}^{N} = f_{transition\_retro}(m_0, S_i, \beta_i, L_{\text{retro}})$$

This function maps a single set of inputs:

- m_0: The atom-mapped target product molecule canoncialized SMILES.
- S_i: The set of disconnection point atom indices.

- β_i: The reaction name, serving as the actionable attribute a_j.
- L_{retro}: The context library, containing examples of the reaction β_i.

to a set of N distinct tuples:

- R_k: The k-th predicted set of reactant molecules $\{r_1, r_2, \ldots, r_n\}$.
- γ_k : The specific chemical validity assessment (stability, chemoselectivity, stereochemical consistency) for the transformation leading to R_k.
- ω_k: The specific chemical rationale that justifies the validity of the k-th outcome.

The transition function $f_{transition\_retro}$ is defined by prompt $r_{transition}$ (see Prompt 2), which emulates a chemist's reasoning and avoids explicit reaction rules. Beyond reactant prediction, the model can also generalize transformations by abstracting a reaction template R_t, which is flagged accordingly. This template can handle complex cases, such as multiple atoms being viable for reaction side or added reagents, thereby preventing exhaustive iteration. We evaluate performance by comparing the predicted reactant sets, $R_{\text{pred}} = \{R_1, \ldots, R_N\}$, against the ground-truth reactants, R_{gt}. As multiple reactant sets can be chemically valid, our goal is to assess the model's ability to recover the known, ground-truth transformation without ranking. The following metrics are calculated per-prediction and averaged across the dataset.

1. Template Accuracy: measures if any predicted reactant template set $R_t \in R_{\text{pred}}$ correctly identifies the core structure of the ground-truth reactants R_{gt}. A prediction is considered a match if for every ground-truth reactant $r_{gt} \in R_{gt}$ there is a corresponding predicted reactant template $r_t \in R_t$ sharing at least 75% of its atoms and having a direct substructure match.
2. Reactant Accuracy: measures if any predicted reactant set R_k is an exact, non-template match for the ground-truth set R_{gt}.
3. Combined Accuracy: measures if a prediction meets either the Template or Reactant Accuracy criterion.

2.4 Experimental Setup

We evaluate the Position ($f_{position\_retro}$) and Transition ($f_{transition\_retro}$) models across a diverse set of LLMs to assess the scaling of reasoning capabilities. Our selection includes various open-source models (Qwen3-2507 4B, 30B, 235B Yang et al. (2025), DeepSeek-R1-0528 Guo et al. (2025)), several closed-source models (Gemini 2.5 Flash/Pro Comanici et al. (2025), Claude Sonnet 4 Anthropic (2025), GPT5 OpenAI (2025)), and a chemistry-specialized model, ether0 Narayanan et al. (2025). For efficiency, the largest open-source models were quantized for inference on an 8x H100 DGX node and used default inference parameters (see Table 2).

We use two public reaction datasets: USPTO50k Lowe (2012); Schneider et al. (2016) and PaRoutes Genheden & Bjerrum (2022). For USPTO50k ($n \approx 5 \times 10^4$), we use an adjusted version that corrects a known atom-mapping bias Somnath et al. (2021). For PaRoutes ($n \approx 1 \times 10^6$), we use the provided data splits Torren-Peraire et al. (2024). For all datasets, we preprocess the data to generate structural labels (S_i), reaction names (β_i) and reaction ontology (O). The labels (S_i) define the reaction center by annotating atoms of bonds that are broken, formed, or changed in type from the product's perspective. We prioritize changes in connectivity (bonds breaking or forming) over bond type changes, where the atom structure itself remains unchanged, unless no connectivity change occurs. The reaction names (β_i) and their reaction classes are extracted using the open-source rxn-insight package Dobbelaere et al. (2024), allowing the release of our labeled data. The ontology (O) is constructed from unique reaction names (β_i) in the respective training data. To mitigate the skewed distribution of reaction names in the USPTO50k test set ($n = 5 \times 10^3$) and prevent redundant evaluation, we create a subsampled version, USPTO50k-LLM (see Fig. 5). This 541-point evaluation set contains up to five examples per unique reaction name, preserving the original proportion of unclassified reactions. Unless specified otherwise, we use this set with a reaction ontology ($n = 136$) derived from the USPTO50k training data.

3 RESULTS

3.1 POSITION MODEL

Our analysis of structural chemical reasoning shows performance scales with model size, with large closed-source models such as the top-performing Gemini 2.5 Pro required for the best results (Fig. 2). We evaluated models on four tasks of increasing difficulty: partial position match, maximizing Jaccard overlap, exact position match, and correct reaction prediction given a partial match. A consistent pattern emerged, where performance increased with the size of the model. For instance, partial match scores jumped from 73% for 4B models to 87% for 235B+ models. This trend held across all tasks, with the performance gap becoming most stark on the reaction prediction task, where smaller models scored just 4%. In contrast, only the largest proprietary models achieved a moderate success rate of 40-47%, showing a trade-off between higher accuracy and lower prediction efficiency (i.e., more predictions per success; see Table 3).

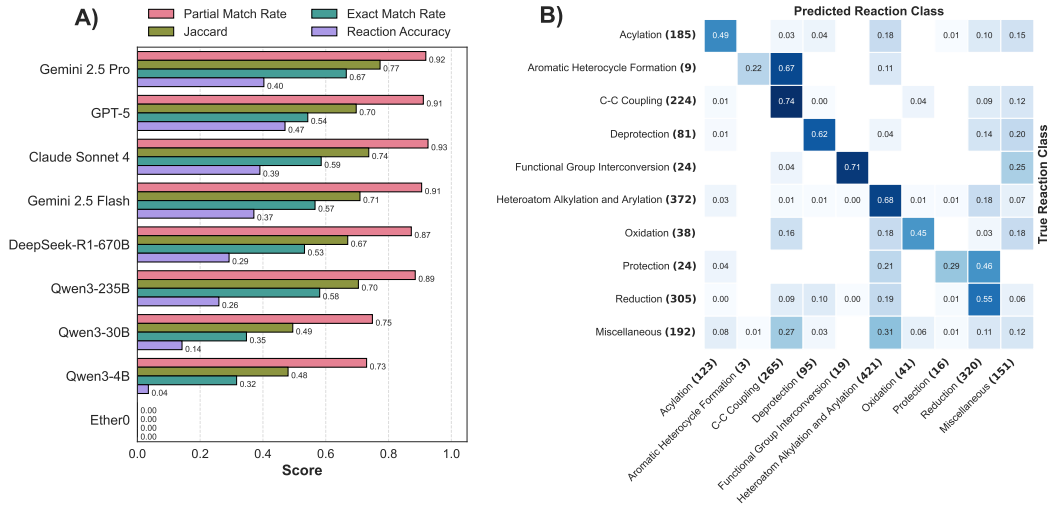


Figure 2: **A)** Position model performance on USPTO-LLM. The plot compares various foundation models on the task of reaction position prediction, measured by four evaluation metrics: achieving a partial positional match, maximizing the Jaccard metric, identifying the exact position, and predicting the correct reaction (conditional on a partial match). **B)** Confusion matrix of predicted versus ground-truth reaction classes for the Gemini 2.5 Pro model on USPTO-LLM. The analysis is conditional, including only predictions where the model successfully identified at least a partial positional match. For this visualization, reactions outside the defined reaction ontology were excluded. The matrix was generated using the original class-to-name mappings from the ground-truth data, with any unassigned reactions grouped into the 'Miscellaneous' category.

Three models warrant a specific discussion. First, the ether0 model, a Mistral-24B variant fine-tuned for chemistry, fails to produce any valid predictions, generating neither valid outputs nor chemically valid positions, unlike other models that fail only occasionally (see Table 3). This total failure suggests that its specialized training, which utilizes chemistry reasoning traces and GRPO on chemical tasks, hindered generalizability to our problem. Second, an ablation of Qwen-235B-Instruct reveals a trade-off with its thinking counterpart. Despite a comparable partial match score, the instruct model showed poor prediction efficiency, generating far more candidate positions, and was only half as effective at identifying the correct reaction (see Table 3), highlighting the importance of CoT reasoning. Interestingly, this pattern does not appear for Gemini 2.5 Flash, where its thinking and non-thinking versions perform comparably with high reaction accuracy and low prediction efficiency.

Our problem involves a one-to-many relationship in which a chemical position can have multiple valid reactions. To evaluate one of the best performing models, Gemini 2.5 Pro, we mapped its predictions to broader reaction classes using the reaction class mapping from rxn-insight on the ground truth data (see Fig. 2). The model often suggests alternative reactions from the correct class rather than predicting a reaction from a different class. However, some exceptions represented chemically

plausible alternative strategies: for 'Aromatic Heterocycle Formation', the model often predicted 'C-C couplings', and for 'Protection' reactions, it suggested 'Reductions'. The 'Heteroatom Alkylation and Arylation' class was a notable outlier, being proposed for most other categories except 'FGI' and 'C-C couplings'. This predictive pattern of staying within-class and these specific exceptions also holds at the individual reaction-name level (see Fig. 6).

3.2 TRANSITION MODEL

We evaluated various LLMs on their ability to predict ground-truth transformations using the reaction's position, name, and up to five examples (Figure 3). Model performance scales logarithmically with size before plateauing at the scale of Deepseek-R1. Gemini 2.5 Pro is the top performer, excelling both at direct reactant prediction ("Reactant"; see example Fig. 11) and in combination with a reaction template ("Combined"). This template generation ("Template"; see example Fig. 12), which is a proxy for chemical understanding, is strongest in proprietary models, such as GPT-5 and Gemini 2.5 Pro (44% accuracy). In contrast, Deepseek-R1 performs worse than its smaller open-source peers in template prediction, while ether0 fails again at this task.

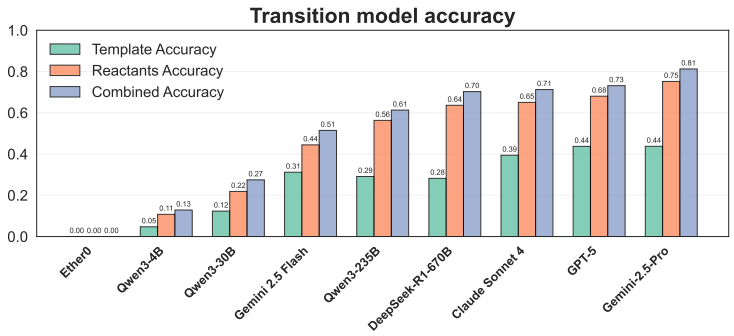


Figure 3: Transition Model Performance on USPTO-LLM. The plot evaluates various LLMs on their ability to predict chemical transformations. Accuracy is measured using three metrics: direct reactant prediction ('Reactants'), valid template generation ('Template'), and a combined approach where either is considered a success ('Combined').

In a first ablation study, our results reveal the critical importance of a defined reaction name to act as a *chemical anchor* (Appendix Fig. 7). Performance dropped by approx. 50% for unknown reactions in a zero-shot setting (no examples provided) compared to known ones in a few-shot setting (up to five examples). The decline was particularly severe for the prediction of direct reactants, with accuracy falling from approximately 75% to 30%. In a second ablation study on Gemini 2.5 Pro, we further isolate the contributions of prompt detail versus few-shot examples on overall ("combined") performance (Appendix Figure 8). Although the model achieved (52%) baseline accuracy from a minimal prompt, and the detailed prompts offered some improvement through the reaction template (59%), the inclusion of examples was the dominant factor (69%); a simple prompt with examples was much more effective than a detailed prompt without them. The best performance required both (81%). Finally, CoT reasoning improves reactant and combined accuracy, but it underperforms non-reasoning models on reaction template prediction, at the cost of lower prediction efficiency (compare Qwen3-235B Table 4).

Analyzing LLM failure modes reveals two distinct error types (Appendix Fig. 9). First, reaction class-specific performance variations among the top-performing models indicate that no single model is universally superior, suggesting solutions such as multi-model ensembles or best-of-n sampling. Second, all models consistently fail on a small set of reaction classes with few data points (e.g., Wohl-Ziegler bromination). This systemic failure likely stems from data deficiencies, such as incorrect labeling, poor examples that make the task ill-posed, rather than fundamental mechanistic reasoning challenges for current LLM architectures.

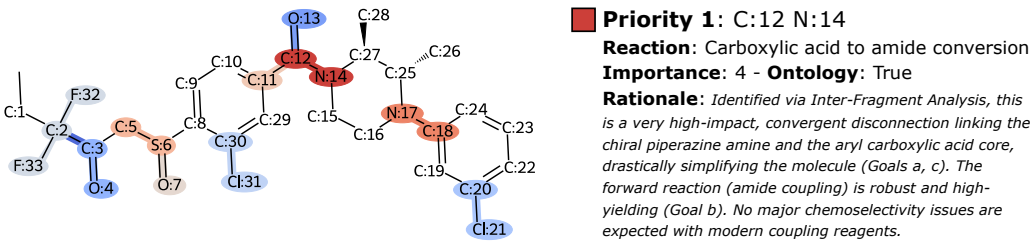


Figure 4: Zero-shot position model prediction for compound LEI-515 Jiang et al. (2023) using the PaRoutes reaction ontology highlighting reaction priority 1. See Table 5 for all priorities (1-14).

4 APPLICATION

While LLMs demonstrated strong performance on USPTO50k, such academic tests risk data contamination for models pre-trained on vast data corpora. To conduct a more rigorous, real-world validation, we evaluated our approach on five molecules that were previously synthesized and published in high-impact journals (see Fig. 10), for which we were able to discuss the experimental procedures with the respective lab chemists. Although this small sample size prevents broad statistical generalization, the case study provides a crucial assessment of the model's practical capabilities and limitations. For this evaluation, we used one of our top-performing LLMs (Gemini 2.5 Pro) with the PaRoutes reaction ontology (n=335) and annotated atom-maps by sequentially counting the atoms in a canonicalized SMILES. Our position model first proposed potential disconnection points, which the respective lab chemist of the molecule then curated for chemical relevance and to avoid redundancy for the transition model evaluation (an example for LEI-515 is provided in Appendix 5). This process yielded 63 distinct position predictions for assessment and 19 selected positions with a total of 98 transitions. Afterwards, the chemist assessed these predictions against predefined questions, and we calculated accuracy as the percentage of correct model responses (compare Table 1).

Table 1: Questions for chemists with regard to the Position model (P) and Transition Model (T). n indicates here the overall number of data points and accuracy (Acc.), as well as the percentage of correct predictions. Actionable refers here to non-template and not to chemically invalid predicted reactant sets from the model. We provide a full overview in the appendix (see Table 6 & 7)

Question	n	Acc.
P1: Disconnection position chemically plausible?	63	90.5
P2: Reaction correct for the proposed disconnection position?	63	85.7
P3: Chemical reasoning correct for the position and reaction?	63	73.0
P4: Given all the information, could this reaction realistically work in the lab?	63	77.8
P5: Specific reaction successfully performed in the lab for the molecule?	63	25.4
P6: Strategically important disconnection predictions missing for the molecule?	5	80.0
T1: Given a predicted reaction template, does it capture the underlying reaction?	16	81.3
T2: Given a predicted reaction template, is the chemical reasoning correct?	16	87.5
T3: Among the reactant predictions, is there at least one chemically correct set?	19	89.5
T4: Given the correct set of reactants, is the chemical reasoning also correct?	19	89.5
T5: Given the reaction was used in the lab, are the predicted reactants the same?	15	73.3
T6: Given that the reactants are flagged 'chemically invalid', is the reasoning correct?	7	100
T7: What % of all the actionable suggested reactants are chemically correct?	98	74.5

The case study results were highly encouraging. The model's suggested disconnection points (90.5%) and associated reaction names (85.7%) were overwhelmingly judged as chemically plausible, with the latter often providing non-obvious alternatives to our expert chemists. While the

accuracy for chemical reasoning was lower (73.0%), a majority of all suggestions (77.8%) were deemed applicable in a laboratory setting. Notably, the model rediscovered 25.4% of the experimentally validated disconnections. This figure is lower because the model often proposes multiple valid reactions for a single position, where only one would be used in practice. However, the system has limitations. For four of the five molecules evaluated, the model missed disconnections anticipated by our chemists. It might, for example, propose a feasible reaction (e.g., Buchwald-Hartwig coupling) where an expert would prefer an alternative (e.g., an S_NAr reaction). Our analysis indicates that errors typically originate from the LLM's misinterpretation of the molecular structure (compare, for instance, the misidentified Cl position in Table 5, position 10). This initial error then propagates through the prediction, ultimately leading to an incorrect suggestion for the position, reaction or reasoning.

A key strength of the position model is its ability to provide a comprehensive set of plausible disconnections for an entire synthetic route, not just a single retrosynthetic step. Our chemists considered these predictions valid if the proposed disconnection could occur at any stage of the synthesis route. This holistic approach has two important consequences: First, the generated positions constrain the search space for a synthesis planning algorithm (e.g., Hassen et al. (2025)), streamlining the identification of an optimal reaction sequence. Second, these predictions highlight vectors for molecular modification, proving invaluable for guiding and accelerating medicinal chemistry campaigns by providing a strategic blueprint for replacing molecular cores or side-chains.

The transition model also demonstrated strong performance. It achieved 81.3% accuracy for predicting reaction templates and 87.5% for the associated reasoning, although chemists noted it worked mainly for standard reactions and is less reliable for complex ones (see Fig. 12). In 89.5% of cases, the model generated at least one chemically valid reactant set with sound reasoning (see Fig. 11), a reasoning quality judged comparable to that of a master's or PhD-level chemist. Furthermore, it successfully identified 73.3% of reactants previously conducted in the lab. A key strength was its perfect (100%) accuracy in identifying non-viable reactions (see Fig. 13), correctly explaining why a proposed reaction would fail (e.g., identifying that a specific atom cannot exist at a given position). This highlights its role as a filter, as it sometimes corrected position model suggestions by proposing more intuitive reactions or filtering out disconnections that were invalid without prerequisite synthesis steps. The model achieved a 74.5% overall accuracy in predicting reactants after excluding predictions that were reaction template-based or flagged as chemically invalid. Failures typically occurred in one of two ways: the model either failed to return any valid reactant set (accounting for 15/29 failures in our evaluation), or it failed due to incorrect SMILES parsing (see Fig. 14), even when the underlying chemical reasoning was correct.

5 Conclusion

We introduce a reasoning framework that leverages the chemical knowledge in general-purpose LLMs to address data scarcity in computational chemistry, requiring no labeled training data or task-specific model training. Our framework anchors chain-of-thought reasoning to the molecular structure by using atom maps in molecular SMILES as chemical anchors and operates in two stages: a one-shot position model identifies relevant molecular fragments and their associated chemical labels or transformations, and an optional position-aware few-shot transition model executes chemical transformations based on provided class examples. Applied to single-step retrosynthesis without task-specific training, our method effectively identifies chemically valid and strategically sound disconnection positions, their corresponding reaction classes, and reactant structures for both academic and expert-validated real-world drug molecules, while providing a chemically grounded, explainable rationale for each prediction. Our work immediately enables new approaches in chemistry, where identified disconnection points define the search space for synthesis planning Westerlund et al. (2025); Kreutter & Reymond (2023) or can be used with a user-defined reaction ontology for robotic or parallel chemistry (e.g., Dombrowski et al. (2022)). More broadly, our framework demonstrates that LLMs can generate realistic synthetic datasets by mapping high-level chemical concepts directly to molecular structures, which in turn can guide applications such as the design of novel, synthetically feasible molecules in de novo drug design. Ultimately, our methodology provides a general blueprint for applying LLM reasoning to various challenges in the molecular sciences where substructure identification is key, establishing atom-anchored LLMs as a powerful and data-efficient addition to the modern drug discovery toolbox.

LARGE LANGUAGE MODELS

Large Language Models (LLMs) were used throughout the creation of this manuscript to improve spelling mistakes, grammar, and the overall reading flow. All LLM suggestions were profusely checked for correctness and refined by the authors of this work. The LLM was not used for any research-related tasks.

REPRODUCIBILITY STATEMENT

The code for *AAL-Chem* can be found on an anonymized repository at https://github.com/AAL-Chem/AAL-Chem. The datasets and raw LLM response files can be found in the DATA/ directory. Figures and tables used in this manuscript can be reproduced via Jupyter notebooks included in the NOTEBOOKS/ directory.

REFERENCES

Mistralai/Mistral-Small-24B-Instruct-2501 · Hugging Face. https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501.

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.

Anthropic. Claude 4 System Card. https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf, 2025.

Marc P. Baggelaar, Pascal J. P. Chameau, Vasudev Kantae, Jessica Hummel, Ku-Lung Hsu, Freek Janssen, Tom Van Der Wel, Marjolein Soethoudt, Hui Deng, Hans Den Dulk, Marco Allarà, Bogdan I. Florea, Vincenzo Di Marzo, Wytse J. Wadman, Chris G. Kruse, Herman S. Overkleeft, Thomas Hankemeier, Taco R. Werkman, Benjamin F. Cravatt, and Mario Van Der Stelt. Highly Selective, Reversible Inhibitor Identified by Comparative Chemoproteomics Modulates Diacylglycerol Lipase Activity in Neurons. *Journal of the American Chemical Society*, 137(27):8851–8857, July 2015. ISSN 0002-7863, 1520-5126. doi: 10.1021/jacs.5b04883.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06792-0.

Andres M. Bran, Theo A. Neukomm, Daniel P. Armstrong, Zlatko Jončev, and Philippe Schwaller. Chemical reasoning in LLMs unlocks strategy-aware synthesis planning and reaction mechanism elucidation, July 2025.

Joseph M. Cavanagh, Kunyang Sun, Andrew Gritsevskiy, Dorian Bagni, Thomas D. Bannister, and Teresa Head-Gordon. SmileyLlama: Modifying Large Language Models for Directed Chemical Space Exploration, September 2024.

Shuan Chen and Yousung Jung. Deep Retrosynthetic Reaction Prediction using Local Reactivity and Global Attention. *JACS Au*, 1(10):1612–1620, 2021. doi: 10.1021/jacsau.1c00246.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

E. J. Corey and Xue-Min Cheng. *The Logic of Chemical Synthesis*. John Wiley & Sons, Ltd, New York, 1989. ISBN 978-0-471-50979-0.

Hui Deng, Sander Kooijman, Adrianus M. C. H. Van Den Nieuwendijk, Daisuke Ogasawara, Tom Van Der Wel, Floris Van Dalen, Marc P. Baggelaar, Freek J. Janssen, Richard J. B. H. N. Van Den Berg, Hans Den Dulk, Benjamin F. Cravatt, Herman S. Overkleeft, Patrick C. N. Rensen, and Mario Van Der Stelt. Triazole Ureas Act as Diacylglycerol Lipase Inhibitors and Prevent Fasting-Induced Refeeding. *Journal of Medicinal Chemistry*, 60(1):428–440, January 2017. ISSN 0022-2623, 1520-4804. doi: 10.1021/acs.jmedchem.6b01482.

Maarten R. Dobbelaere, István Lengyel, Christian V. Stevens, and Kevin M. Van Geem. RxnINSIGHT: Fast chemical reaction analysis using bond-electron matrices. *Journal of Cheminformatics*, 16(1):37, March 2024. ISSN 1758-2946. doi: 10.1186/s13321-024-00834-z.

Amanda W. Dombrowski, Ana L. Aguirre, Anurupa Shrestha, Kathy A. Sarris, and Ying Wang. The Chosen Few: Parallel Library Reaction Methodologies for Drug Discovery. *The Journal of Organic Chemistry*, 87(4):1880–1897, February 2022. ISSN 0022-3263, 1520-6904. doi: 10.1021/acs.joc.1c01427.

Samuel Genheden and Esben Bjerrum. PaRoutes: Towards a framework for benchmarking retrosynthesis route predictions. *Digital Discovery*, 1(4):527–539, 2022. doi: 10.1039/D2DD00015F.

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, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2023. Curran Associates Inc.

Alan Kai Hassen, Helen Lai, Samuel Genheden, Mike Preuss, and Djork-Arné Clevert. Synthesis Planning in Reaction Space: A Study on Success, Robustness and Diversity, June 2025.

Ilia Igashov, Arne Schneuing, Marwin Segler, Michael Bronstein, and Bruno Correia. Retrobridge: Modeling retrosynthesis with markov bridges. 2024.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: A pretrained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3 (1):015022, March 2022. ISSN 2632-2153. doi: 10.1088/2632-2153/ac3ffb.

Ming Jiang, Mirjam C. W. Huizenga, Jonah L. Wirt, Janos Paloczi, Avand Amedi, Richard J. B. H. N. Van Den Berg, Joerg Benz, Ludovic Collin, Hui Deng, Xinyu Di, Wouter F. Driever, Bogdan I. Florea, Uwe Grether, Antonius P. A. Janssen, Thomas Hankemeier, Laura H. Heitman, Tsang-Wai Lam, Florian Mohr, Anto Pavlovic, Iris Ruf, Helma Van Den Hurk, Anna F. Stevens, Daan Van Der Vliet, Tom Van Der Wel, Matthias B. Wittwer, Constant A. A. Van Boeckel, Pal Pacher, Andrea G. Hohmann, and Mario Van Der Stelt. A monoacylglycerol lipase inhibitor showing therapeutic efficacy in mice without central side effects or dependence. *Nature Communications*, 14(1):8039, December 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-43606-3.

Seongmin Kim, Yousung Jung, and Joshua Schrier. Large Language Models for Inorganic Synthesis Predictions, June 2024.

David Kreutter and Jean-Louis Reymond. Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search. *Chemical Science*, 14(36): 9959–9969, 2023. ISSN 2041-6520, 2041-6539. doi: 10.1039/D3SC01604H.

Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. Beyond Chemical QA: Evaluating LLM's Chemical Reasoning with Modular Chemical Operations, June 2025.

Xiaoting Li, Hao Chang, Jara Bouma, Laura V. De Paus, Partha Mukhopadhyay, Janos Paloczi, Mohammed Mustafa, Cas Van Der Horst, Sanjay Sunil Kumar, Lijie Wu, Yanan Yu, Richard J. B. H. N. Van Den Berg, Antonius P. A. Janssen, Aron Lichtman, Zhi-Jie Liu, Pal Pacher, Mario Van Der Stelt, Laura H. Heitman, and Tian Hua. Structural basis of selective cannabinoid CB2 receptor activation. *Nature Communications*, 14(1):1447, March 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-37112-9.

Daniel Mark Lowe. *Extraction of Chemical Structures and Reactions from the Literature*. Thesis, University of Cambridge, 2012.

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*, 6(5):525–535, May 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00832-8.

Muhammad Arslan Masood, Samuel Kaski, and Tianyu Cui. Molecular property prediction using pretrained-BERT and Bayesian active learning: A data-efficient approach to drug design. *Journal of Cheminformatics*, 17(1):58, April 2025. ISSN 1758-2946. doi: 10.1186/s13321-025-00986-6.

Elliot D. Mock, Mohammed Mustafa, Ozge Gunduz-Cinar, Resat Cinar, Gavin N. Petrie, Vasudev Kantae, Xinyu Di, Daisuke Ogasawara, Zoltan V. Varga, Janos Paloczi, Cristina Miliano, Giulia Donvito, Annelot C. M. Van Esbroeck, Anouk M. F. Van Der Gracht, Ioli Kotsogianni, Joshua K. Park, Andrea Martella, Tom Van Der Wel, Marjolein Soethoudt, Ming Jiang, Tiemen J. Wendel, Antonius P. A. Janssen, Alexander T. Bakker, Colleen M. Donovan, Laura I. Castillo, Bogdan I. Florea, Jesse Wat, Helma Van Den Hurk, Matthias Wittwer, Uwe Grether, Andrew Holmes, Constant A. A. Van Boeckel, Thomas Hankemeier, Benjamin F. Cravatt, Matthew W. Buczynski, Matthew N. Hill, Pal Pacher, Aron H. Lichtman, and Mario Van Der Stelt. Discovery of a NAPE-PLD inhibitor that modulates emotional behavior in mice. *Nature Chemical Biology*, 16 (6):667–675, June 2020. ISSN 1552-4450, 1552-4469. doi: 10.1038/s41589-020-0528-7.

Siddharth M. Narayanan, James D. Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Samuel G. Rodriques, and Andrew D. White. Training a Scientific Reasoning Model for Chemistry, June 2025.

Phuong Nguyen-Van, Long Nguyen Thanh, Ha Hoang Manh, Ha Anh Pham Thi, Thanh Le Nguyen, and Viet Anh Nguyen. Adapting Language Models for Retrosynthesis Prediction, July 2024.

OpenAI. GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025.

Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Yejin Choi, Jiawei Han, and Lianhui Qin. Structured Chemistry Reasoning with Large Language Models, February 2024.

Chen Qian, Huayi Tang, Zhirui Yang, Hong Liang, and Yong Liu. Can Large Language Models Empower Molecular Property Prediction?, July 2023.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties, December 2022.

Shaghayegh Sadeghi, Alan Bui, Ali Forooghi, Jianguo Lu, and Alioune Ngom. Can large language models understand molecules? *BMC Bioinformatics*, 25(1):225, June 2024. ISSN 1471-2105. doi: 10.1186/s12859-024-05847-x.

Nadine Schneider, Nikolaus Stiefl, and Gregory A. Landrum. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *Journal of Chemical Information and Modeling*, 56(12): 2336–2346, 2016. doi: 10.1021/acs.jcim.6b00564.

Marwin H.S. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610, 2018. ISSN 1476-4687. doi: 10.1038/nature25978.

Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, and Regina Barzilay. Learning Graph Models for Retrosynthesis Prediction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9405–9415. Curran Associates, Inc., 2021.

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. doi: 10.1038/s41467-020-19266-y.

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, July 2023. ISSN 2374-7943, 2374-7951. doi: 10.1021/acscentsci.3c00372.

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(3):558–572, 2024. ISSN 2635-098X. doi: 10.1039/D3DD00252G.

Haorui Wang, Jeff Guo, Lingkai Kong, Rampi Ramprasad, Philippe Schwaller, Yuanqi Du, and Chao Zhang. LLM-Augmented Chemical Synthesis and Design Decision Programs, May 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023.

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, February 1988. ISSN 0095-2338, 1520-5142. doi: 10.1021/ci00057a005.

David Weininger, Arthur Weininger, and Joseph L. Weininger. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, May 1989. ISSN 0095-2338, 1520-5142. doi: 10.1021/ci00062a008.

Annie M. Westerlund, Lakshidaa Saigiridharan, and Samuel Genheden. Human-guided synthesis planning *via* prompting. *Chemical Science*, 16(32):14655–14667, 2025. ISSN 2041-6520, 2041-6539. doi: 10.1039/D5SC00927H.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, May 2025.

Yifei Yang, Runhan Shi, Zuchao Li, Shu Jiang, Bao-Liang Lu, Yang Yang, and Hai Zhao. BatGPT-Chem: A Foundation Large Model For Retrosynthesis Prediction, August 2024.

Weihe Zhong, Ziduo Yang, and Calvin Yu-Chian Chen. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications*, 14(1): 3009, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38851-5.

A APPENDIX

A.1 EXPERIMENTAL SETUP

Table 2: A summary of the Large Language Models (LLMs) evaluated in this work. The table specifies whether the model is open-source, its status as a reasoning-optimized ("Thinking") variant, and its thinking budget allocation (in number of tokens) for closed-source models along with other parameters.

Source	Model Name	Thinking model	Open-Source	Model quantization	Max output length	Thinking budget
Yang et al. (2025)	Qwen3-4B-Thinking-2507	yes	yes		32768	-
Narayanan et al. (2025)	ether0 (24B)	yes	yes		32768	-
Yang et al. (2025)	Qwen-3-30B-A3B-Thinking-2507	yes	yes	8bit	32768	-
Yang et al. (2025)	Qwen3-235B-A22B-Instruct-2507-FP8	no	yes	8bit	32768	-
Yang et al. (2025)	Qwen3-235B-A22B-Thinking-2507-FP8	yes	yes	8bit	32768	-
Guo et al. (2025)	RedHat-DeepSeek-R1-0528-w4a16 (670B)	yes	yes	4bit	32768	-
Comanici et al. (2025)	Gemini 2.5 Flash	yes	no	API	65536	30000
Comanici et al. (2025)	Gemini 2.5 Pro	yes	no	API	65536	30000
Anthropic (2025)	Claude Sonnet 4	yes	no	API	64000	30000
OpenAI (2025)	GPT5	yes	no	API	128000	'High'

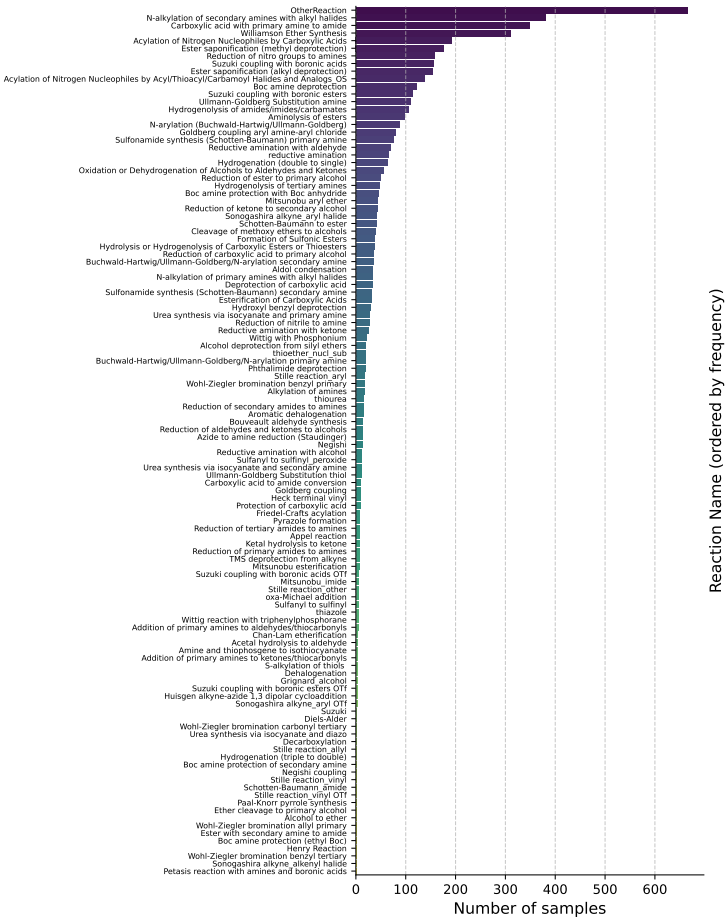


Figure 5: Distribution of reaction names in the USPTO-50k test set. From this dataset, we created a balanced subsample (USPTO-LLM) for evaluation by selecting up to five examples per named reaction class, while maintaining the original proportion of the 'otherReaction' class.

A.2 Position Model

Table 3: A comprehensive comparison of various models based on several key performance metrics. The table highlights the average number of predictions, partial and exact match percentages, reaction accuracy, and the total number of successes and failures for each model. The best performance in each column is highlighted in bold.

Model	Avg. number of predictions	Partial match (%)	Exact match (%)	Reaction acc. (%)	Total predictions	Failed predictions
Ether0	0.0	0.0	0.0	0.0	0	541
Qwen3-4B	4.0	73.01	31.61	3.51	541	0
Qwen3-30B	3.8	74.86	34.75	14.23	541	0
Gemini 2.5 Flash	15.3	90.54	56.59	37.11	539	2
Gemini 2.5 Flash (thinking)	**16.3**	91.84	61.6	35.81	539	2
Qwen3-235B-thinking	5.9	88.5	58.07	25.97	539	2
Qwen3-235B-instruct	9.6	86.67	49.44	13.33	540	1
DeepSeek-R1-670B	7.3	87.25	53.23	29.21	541	0
Claude Sonnet 4	10.0	**92.57**	58.55	39.03	538	3
GPT-5	15.1	91.08	54.28	**47.03**	538	3
Gemini 2.5 Pro	11.1	91.87	**66.54**	40.3	541	0

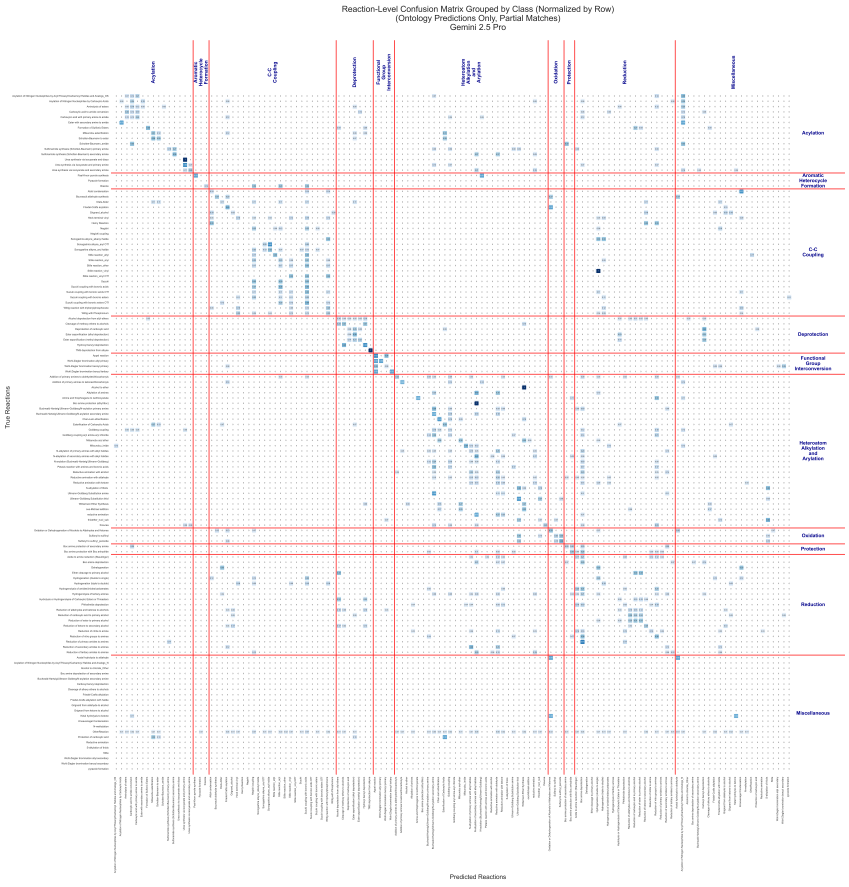


Figure 6: Confusion matrix of predicted versus ground-truth reaction names for the Gemini 2.5 Pro model. The analysis is conditional, including only predictions where the model successfully identified at least a partial positional match. For this visualization, reactions outside the defined reaction ontology were excluded. The matrix was generated using the original class-to-name mappings from the ground-truth data, with any unassigned reactions grouped into the 'Miscellaneous' category.

A.3 TRANSITION MODEL

This section covers additional results on the transition model (reactant prediction).

Table 4: A comparison of model performance on the transition task (reactant prediction). This table presents the total successful predictions, along with accuracy scores for reactants, templates, and combined category. The best performance in each column is highlighted in bold.

Model	Avg. number of predictions	Reactants accuracy	Template accuracy	Combined accuracy	Total predictions	Failed predictions
Ether0	0.0	0.0	0.0	0.0	0.0	541
Qwen3-4B	3.0	0.11	0.05	0.13	529	12
Qwen3-30B	3.6	0.22	0.12	0.27	535	6
Gemini 2.5 Flash	4.4	0.44	0.31	0.51	513	28
Qwen3-235B-thinking	4.4	0.56	0.29	0.61	522	19
Qwen3-235B-instruct	6.6	0.40	0.39	0.48	537	4
DeepSeek-R1-670B	4.4	0.64	0.28	0.70	528	13
Claude Sonnet 4	5.0	0.65	0.39	0.71	515	26
GPT-5	10.4	0.68	**0.44**	0.73	510	31
Gemini 2.5 Pro	5.7	**0.75**	**0.44**	**0.81**	512	29

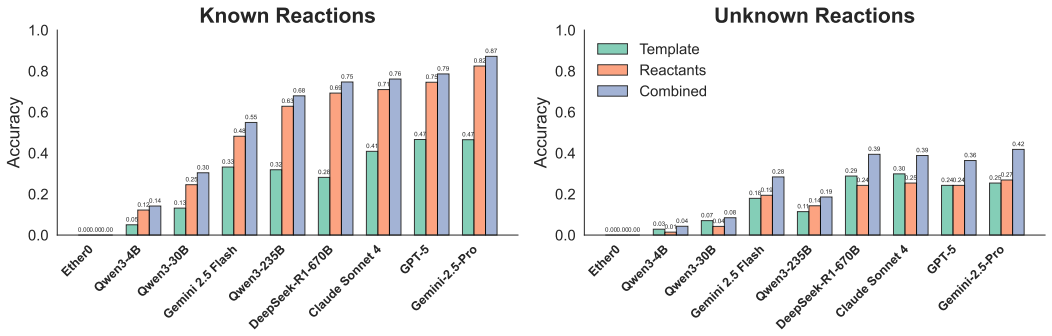


Figure 7: Performance difference between known and unknown reaction names. For unknown reactions, no equivalent name reaction examples within the *USPTO50k* training dataset are provided, illustrating the importance of the reaction name as a chemical anchor for retrieving reaction examples and chemical reasoning.

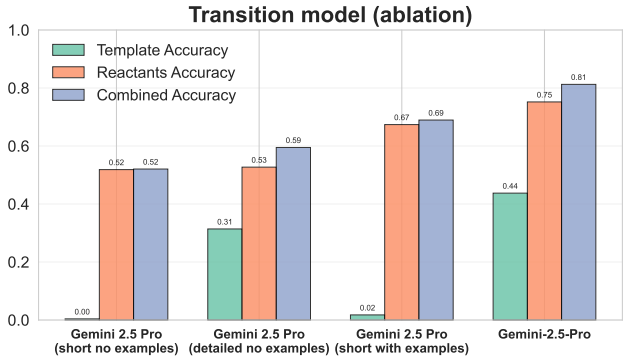


Figure 8: An ablation study on the impact of prompt instruction detail and the inclusion of in-context examples on the performance of the Gemini 2.5 Pro transition model. We evaluate four settings: 1) a simple prompt without examples (see Appendix 3); 2) a detailed prompt without examples (see Appendix 2); 3) a simple prompt with examples; and 4) a detailed prompt with examples.

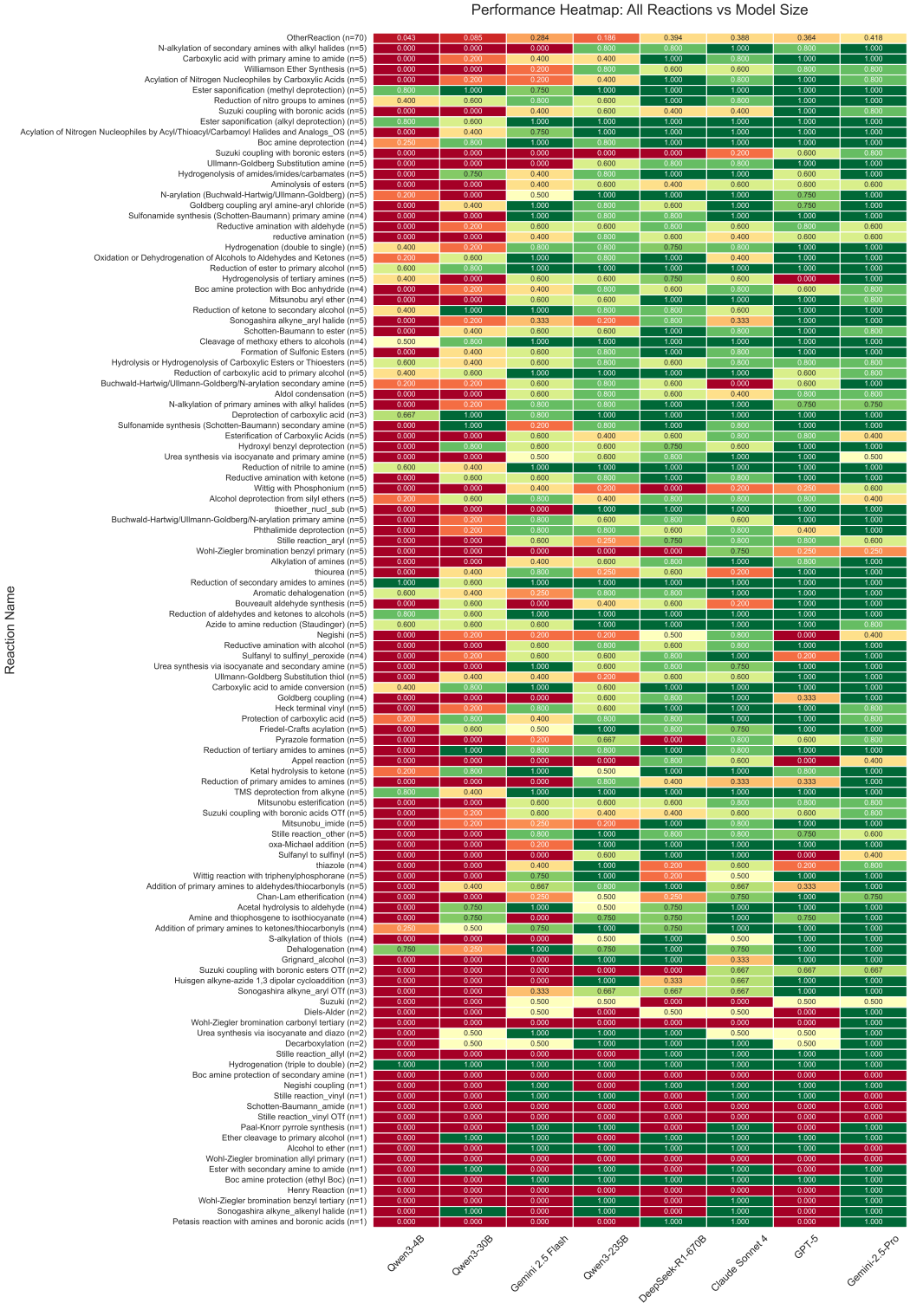


Figure 9: Confusion Matrix Highlighting the performance of different Transition Models on respective reaction name classes using either template or reactant accuracy. The reactions are sorted by the number of reaction examples available in the set (high-to-low).

A.4 Application Examples

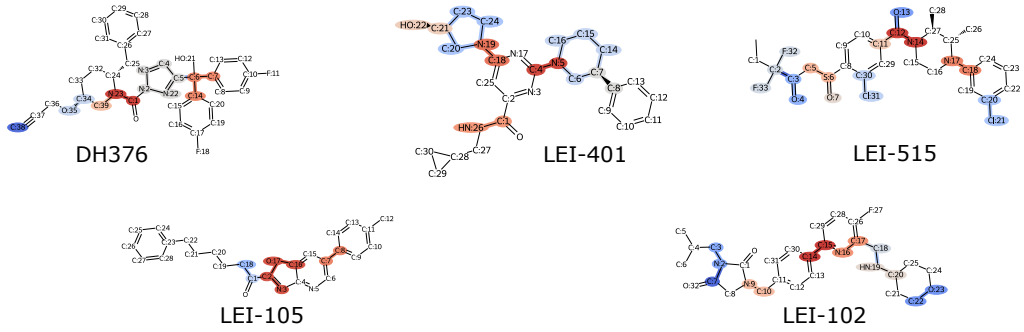


Figure 10: Five real-world drug disovery molecules used in our case study: DH376 Deng et al. (2017), LEI-102 Li et al. (2023), LEI-105 Baggelaar et al. (2015), LEI-401 Mock et al. (2020), LEI-515 Jiang et al. (2023)

Table 5: Predicted Disconnection Sites for LEI_515 Jiang et al. (2023). Header descriptions are as follows: **Prio.**: Priority Ranking of the Disconnections; **Position**: The position where the disconnection is; **Reaction**: The forward reaction; **Ontology**: If the reaction is in the reaction ontology for which examples are available; **Imp.**: Retrosynthesis Importance, alignment with retrosynthesis goals; **Rationale**: The chemical rationale.

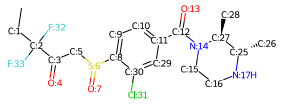
Prio.	Position	Reaction	Ontology	Imp.	Rationale
1	C:12 N:14	Carboxylic acid to amide conversion	Yes	4	Identified via Inter-Fragment Analysis, this is a very high-impact, convergent disconnection linking the chiral piperazine amine and the aryl carboxylic acid core, drastically simplifying the molecule (Goals a, c). The forward reaction (amide coupling) is robust and high-yielding (Goal b). No major chemoselectivity issues are expected with modern coupling reagents.
2	C:12 N:14	Acylation of Nitrogen Nucleophiles by Acyl/Thioacyl/Carbamoyl Halides and Analogs_N	Yes	4	Identified via Inter-Fragment Analysis, this is a very high-impact, convergent disconnection that simplifies the molecule into two key fragments (Goals a, c). Activation of the carboxylic acid as an acyl chloride is a classic, robust method for amide formation (Goal b), though it may require protection of other nucleophilic sites.
3	N:17 c:18	Buchwald-Hartwig/Ullmann-Goldberg/N-arylation secondary amine	Yes	4	Identified via Inter-Fragment Analysis, this key convergent C-N bond formation builds the N-arylpiperazine motif (Goal c). The forward reaction, a Buchwald-Hartwig amination, is a powerful and reliable method for this transformation, simplifying the chiral amine precursor (Goals a, b). Potential for competitive reaction at N:14 necessitates a protecting group strategy.

Prio.	Position	Reaction	Ontology	Imp.	Rationale
4	N:17 c:18	N-arylation (Buchwald-Hartwig/Ullmann-Goldberg)	Yes	4	Identified via Inter-Fragment Analysis, this strategic disconnection simplifies the chiral piperazine component by removing the aryl group (Goal a). The forward N-arylation is a powerful and reliable reaction for constructing this key bond (Goals b, c). Regioselectivity between the two piperazine nitrogens is a key issue that must be controlled, likely with a protecting group on N:14.
5	C:5 S:6	S-alkylation of thiols	Yes	3	Identified via Strategic Bond Analysis. This disconnects the beta-keto sulfoxide side chain from the aryl core (Goal c). The forward synthesis involves S-alkylation of the corresponding thiophenol with an alpha-halo ketone precursor, followed by oxidation. This is a robust way to build this key C-S bond, but the electrophile itself requires separate synthesis (Goals a, b).
6	c:11 C:12	Grignard with CO2 to carboxylic acid	Yes	3	Identified via FGI Analysis. Strategic C-C bond formation to install the carboxyl group on the central aromatic ring (Goal c). The forward synthesis via carboxylation of an organometallic (e.g., Grignard) from an aryl halide is classic and reliable (Goals a, b). Chemoselectivity is a concern, as formation of the organometallic must be compatible with the sulfide/sulfoxide group.
7	S:6 O:7	Sulfanyl to sulfinyl_H2O2	Yes	2	Identified via FGI Analysis. Standard FGI to install the sulfoxide from a more stable and easier to handle sulfide precursor. The oxidation can be performed late-stage, but requires careful control of conditions to prevent over-oxidation to the sulfone, which presents a chemoselectivity challenge (Goal b).
8	C:2 F:32 F:33	Fluorination	Yes	2	Identified via FGI Analysis. This disconnection corresponds to a forward reaction installing the difluoro moiety. Electrophilic fluorination of the beta-keto sulfoxide enolate is a viable route (Goal c). Chemoselectivity could be an issue due to multiple acidic protons (at C:5) and potential for mono- vs di-fluorination, requiring kinetic control.
9	c:30 Cl:31	Aromatic chlorination	Yes	2	Identified via FGI Analysis. This FGI installs the chloro substituent via electrophilic aromatic substitution (Goal c). The regioselectivity of the chlorination would be directed by the existing sulfoxide/sulfide and carboxylate/amide groups. Predicting and controlling the outcome relative to other open positions on the ring requires careful consideration of the combined directing effects.

Prio.	Position	Reaction	Ontology	Imp.	Rationale
10	c:20 Cl:21	Aromatic chlorination	Yes	2	Identified via FGI Analysis. This FGI installs the chloro substituent on the N-aryl ring via electrophilic aromatic substitution (Goal c). The reaction would be strongly directed by the activating amine substituent, likely leading to the observed para-chlorination, making this a reliable and predictable transformation (Goal b).
11	C:3 O:4	Oxidation or Dehydrogenation of Alcohols to Aldehydes and Ketones	Yes	2	Identified via FGI analysis. Standard FGI to form the ketone from a secondary alcohol precursor. While many mild oxidation reagents are available, the presence of the easily oxidizable sulfoxide (or its sulfide precursor) on the same molecule presents a major chemoselectivity challenge that must be carefully managed (Goal b).
12	C:12 O:13 N:14	Nitrile to amide	Yes	2	Identified via FGI analysis. This transforms the amide into a nitrile precursor, offering an alternative synthetic route to the central aromatic core (Goal a). A nitrile can be introduced via methods like the Sandmeyer reaction. The forward reaction, partial hydrolysis of the nitrile to the amide, can be challenging to stop without proceeding to the carboxylic acid.
13	N:14	Boc amine deprotection	Yes	1	Identified via Protecting Group Analysis. This is a tactical deprotection step. A protecting group like Boc on N:14 would be crucial in a forward synthesis to ensure regioselective N-arylation at N:17. This step reveals the nucleophilic amine for the final amide coupling and is a common, practical consideration (Goal d).
14	C:2 C:3	Enolate Acylation	No	3	Identified via Strategic Bond Analysis. This strategic C-C bond disconnection breaks down the beta-keto side chain (Goal a). The forward reaction, likely an enolate acylation, is a powerful method for ketone synthesis (Goal c). However, generating and controlling the reactivity and stability of the required difluoroenolate precursor could be challenging.

Transition 3

- **Forward Reaction:** Buchwald-Hartwig/Ullmann-Goldberg/N-arylation secondary amine
- **Is Chemically Valid:** True
- **Is Template-based:** False
- **Reasoning:** This option is highly plausible. The reactants are stable. The greater reactivity of the C-Br bond versus the C-Cl bonds allows for high chemoselectivity under typical Buchwald-Hartwig conditions. The reaction does not affect the existing stereocenters.
- **Reactant 1:** `C[CH2:1][C:2]([C:3](=[O:4])[CH2:5][S:6](=[O:7])[c:8]1[cH:9][cH:10][c:11]([C:12](=[O:13])[N:14]2[CH2:15][CH2:16][NH:17][C@@H:25]([CH3:26])[C@@H:27]2[CH3:28])[cH:29][c:30]1[Cl:31])([F:32])[F:33]`



- **Reactant 2:** `Br[c:18]1[cH:19][c:20]([Cl:21])[cH:22][cH:23][cH:24]1`

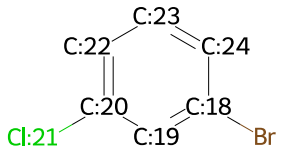
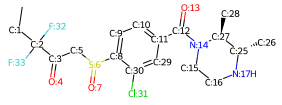


Figure 11: Correct reactant prediction for LEI-515 Jiang et al. (2023) by the Transition model (position priority 3, transition prediction 3).

Transition 1

- **Forward Reaction:** Buchwald-Hartwig/Ullmann-Goldberg/N-arylation secondary amine
- **Is Chemically Valid:** True
- **Is Template-based:** True
- **Reasoning:** This is the general template for an N-arylation reaction. The aryl partner is an activated aryl ring where the wildcard atom [*] represents a suitable leaving group from the class , such as F, Cl, Br, I, OTf, OTs.
- **Reactant 1:** `C[CH2:1][C:2]([C:3](=[O:4])[CH2:5][S:6](=[O:7])[c:8]1[cH:9][cH:10][c:11]([C:12](=[O:13])[N:14]2[CH2:15][CH2:16][NH:17][C@@H:25]([CH3:26])[C@@H:27]2[CH3:28])[cH:29][c:30]1[Cl:31])([F:32])[F:33]`



- **Reactant 2:** `*[c:18]1[cH:19][c:20]([Cl:21])[cH:22][cH:23][cH:24]1`

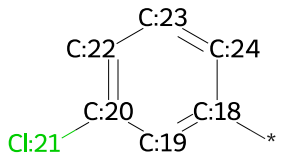


Figure 12: Correct reactant template prediction for LEI-515 Jiang et al. (2023) by the Transition model (position priority 3, transition prediction 1).

Transition 1

- **Forward Reaction:** Boc amine deprotection
- **Is Chemically Valid:** False
- **Is Template-based:** False
- **Reasoning:** This permutation is invalid. The specified reaction center, N:14, is a tertiary amide nitrogen within the product structure, not a primary or secondary amine. The forward reaction 'Boc amine deprotection' implies the product is a free amine and the reactant is a Boc-protected amine (a neutral carbamate). Applying the retrosynthetic protection step to a tertiary amide nitrogen is not a standard or plausible transformation. It would generate a highly reactive and unstable N-acyl quaternary ammonium salt, which is not a viable precursor for a deprotection reaction. The specified reaction name is inconsistent with the provided product structure at the indicated reaction center.
- **Reactant 1:** `C[CH2:1][C:2]([C:3](=[O:4])[CH2:5][S:6](=[O:7])[c:8]1[cH:9][cH:10][c:11]([C:12](=[O:13])[N+:14]2(CC(=O)OC(C)(C)C)C)[CH2:15][CH2:16][N:17]([c:18]3[cH:19][c:20]([Cl:21])[cH:22][cH:23][cH:24]3)[C@@H:25]([CH3:26])[C@@H:27]2[CH3:28])[cH:29][c:30]1[Cl:31])([F:32])[F:33]`

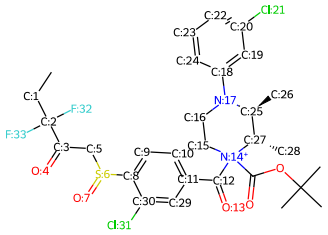
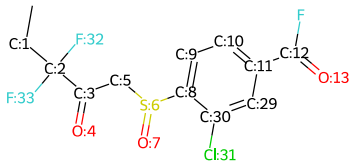


Figure 13: Correctly flagged chemically invalid prediction for LEI-515 Jiang et al. (2023) by the Transition Model (position priority 13, transition prediction 1).

Transition 4

- **Forward Reaction:** Carboxylic acid to amide conversion
- **Is Chemically Valid:** True
- **Is Template-based:** False
- **Reasoning:** This permutation is valid and represents a modern approach to amide synthesis. Acyl fluorides offer a good balance of reactivity and stability, often being more chemoselective and causing less racemization at adjacent stereocenters than the corresponding chlorides. An external base is typically used.
- **Reactant 1:** `[CH3:28][C@@H:27]1[N:14](H)[CH2:15][CH2:16][N:17]([c:18]2[cH:19][c:20]([Cl:21])[cH:22][cH:23][cH:24]2)[C@@H:25]1[CH3:26]` (Could not visualize)
- **Reactant 2:** `C[CH2:1][C:2]([C:3](=[O:4])[CH2:5][S:6](=[O:7])[c:8]1[cH:9][cH:10][c:11]([C:12](=[O:13])F)[cH:29][c:30]1[Cl:31])([F:32])[F:33]`



Reactant 2

Figure 14: Syntactically invalid SMILES prediction for LEI-515 Jiang et al. (2023) by the Transition model (position priority 1, transition prediction 4).

A.5 Application Questionaire

Table 6: Full list of questions for the expert validation study. These are the complete, verbatim questions presented to chemists to benchmark the performance of our framework. The evaluation was split into two parts: assessing the disconnection sites proposed by the Position (P) model and the final reactant structures generated by the Transition (T) model.

Q.	Description
P1	Is the suggested disconnection position chemically plausible (i.e., not violating fundamental principles)?
P2	Is the suggested reaction name a correct label for the proposed disconnection position?
P3	Is the provided chemical reasoning for the suggested disconnection (position and reaction name) scientifically sound?
P4	Considering all the provided information, could this suggested step realistically work in a laboratory setting?
P5	Has this specific transformation actually been performed successfully in practice for the molecule?
P6	Are there any strategically important disconnections that are obviously missing from this prediction?
T1	Given the transition prediction includes a reaction template, does the reaction template capture the overall chemical transformation of the reaction?
T2	Given the transition prediction includes a reaction template, does the chemical reasoning for the reaction template align with the underlying reaction?
T3	Among the reactant predictions, is there at least one that provides a chemically correct set of reactants to form the target product?
T4	If the model predicts a chemically correct set of reactants, is the model's chemical reasoning for that specific set of reactants correct?
T5	If the reaction was conducted in the lab, does the model correctly predict the set of reactants that were used in the lab?
T6	If the model flags one of its own predictions as 'chemically invalid', is its reasoning for that assessment correct?
T7	How many reactants are predicted as chemically valid and are not reaction templates are correct?

Table 7: Detailed reponse data.

ID	P1	P2	P3	P4	P5	P6	T1	T2	T3	T4	T5	T6	T7
DH376 Deng et al. (2017)	12/13	11/13	8/13	11/13	5/13	1	4/4	4/4	4/4	4/4	3/4	2/2	23/31
LEI-102 Li et al. (2023)	14/16	12/16	12/16	14/16	2/16	1	3/3	3/3	4/4	4/4	4/4	3/3	18/18
LEI-105 Baggelaar et al. (2015)	8/9	8/9	8/9	5/9	2/9	1	2/2	2/2	2/2	2/2	1/1	-	11/11
LEI-401 Mock et al. (2020)	11/11	11/11	7/11	11/11	2/11	0	2/3	3/3	3/3	3/3	2/2	1/1	10/15
LEI-515 Jiang et al. (2023)	12/14	12/14	11/14	8/14	5/14	1	2/4	2/4	4/6	4/6	1/4	1/1	11/23
Acc.	90.5	85.7	73.0	77.8	25.4	80.0	81.3	87.5	89.5	89.5	73.3	1	74.5

A.6 PROMPTS

A.6.1 POSITION MODEL

```
**Persona:**
You are an expert chemist specializing in retrosynthetic analysis.

**Primary Goal:**
Your primary goal is to perform a comprehensive retrosynthetic analysis on a given molecule. You will identify all strategically viable disconnection points, rank them according to the provided framework, and format the entire output as a single, valid JSON object.

**Input Schema:**
- product_smiles: The atom-mapped SMILES string of the product molecule.
- reaction_ontology: The provided JSON object containing the reaction ontology.

**Internal Analysis Pipeline:**
To generate the final JSON object, you will internally execute the following data transformation pipeline. The output of each step serves as the direct input for the next, ensuring a dependent, step-by-step analysis.

1.  **Step 1: Identify All Candidate Transformations**
    Process steps A - L sequentially. For each step, you must perform a complete and independent analysis to identify all transformations that fit its description. A finding in one step does not exclude findings in others.
    * **Input:** The `product_smiles`.
    * **Process:**
        * A) **Symmetry Analysis:** First, assess the molecule for any elements of symmetry. If symmetrical fragments exist, identify transformations that could form the molecule by coupling two identical precursors.
        * B) **Fragment Partitioning:** Mentally partition the molecule into its major constituent fragments. The goal is to find disconnections that lead to a **convergent synthesis**.
        * C) **Inter-Fragment Analysis:** Identify the bonds that **connect these major fragments**. These are candidates for strategic coupling reactions.
        * D) **Strategic Bond Analysis:** Within the identified fragments, specifically look for bonds that are adjacent to functional groups, making them chemically activated and strategic targets for disconnection (e.g., bonds alpha/beta to carbonyls, bonds within key functional groups like amides and esters).
        * E) **Intra-Fragment Analysis:** Within each major fragment, identify bonds that could be strategically formed via an **intramolecular (ring-closing) reaction**.
        * F) **Stereochemical Analysis:** Identify all stereocenters. For each one, consider transformations that could set that stereocenter (e.g., asymmetric reactions, chiral pool approach).
        * G) **Rearrangement Analysis:** Look for structural motifs that could be efficiently formed via a powerful **skeletal rearrangement**.
        * H) **FGI Analysis:** For each functional group in the molecule, systematically identify all possible functional groups that are candidates for standard Functional Group Interconversions. This analysis **must** include, but is not limited to:
            * **i. Oxidation/Reduction:** Identify all groups that could be retrosynthetically derived from a different oxidation state.
            * **ii. Non-Redox FGIs:** Identify all non-redox interconversions. This involves analyzing polar carbon-heteroatom bonds within functional groups that are classically disconnected via substitution or hydrolysis-type mechanisms.
        * I) **Protecting Group Analysis:** Analyze for protecting group strategies by proposing protections for sensitive functional groups or deprotections for existing, recognizable protecting groups. Note that a retrosynthethic protection is a forward deprotection reaction and vice versa.
        * J) **Multi-Bond / Multi-Component Analysis:** Analyze the product for structural motifs that could be formed via reactions that form multiple bonds in one step, such as **cycloadditions** (ring-forming reactions between unsaturated systems) or **multi-component reactions** (where 3+ reactants combine in a single operation).
        * K) **Radical Mechanism Analysis:** K) Radical Mechanism Analysis: Analyze the molecule for transformations whose mechanism is best described as proceeding via radical (uncharged, open-shell) intermediates. This involves identifying bonds whose formation or cleavage is characteristic of single-electron processes (homolysis), as distinct from the two-electron processes of polar (ionic) reactions.
        * L) **Novel or Uncategorized Strategies:** If you identify a powerful, chemically sound transformation that does not clearly fit into categories A-K, classify it here.
    * **Output (Internal):** A list of formatted transformation strings representing all identified transformations. Each string must adhere to the format specified for the `"disconnection"` key in the Constraints & Formatting Rules. You MUST return all found disconnections. You are not allowed to leave any found and valid disconnection out.

2.  **Step 2: Assign Candidate Reactions**
    * **Input:** The list of transformation strings from Step 1.
    * **Process:** For each transformation, determine all appropriate forward reaction names. A single transformation may have multiple corresponding reactions.
    * **Output (Internal):** A list of objects, where each object contains a transformation and a list of its assigned `forwardReaction` names.
    * **Example:** `[{ "disconnection": "C:4,C:7", "reactions": ["Suzuki-Miyaura coupling", "Stille coupling"] }]`

3.  **Step 3: Expand and Evaluate Pairs**
    * **Input:** The list of objects from Step 2.
    * **Process:** Expand the input into a flat list by creating a **new, separate entry for each reaction** associated with a transformation. Then, for each of these new entries, apply the Retrosynthetic Analysis Framework to assign a `Retrosynthesis Importance` value and write a concise `rationale`.
    * **Output (Internal):** A flat list of fully populated objects, where each object represents one unique transformation-reaction pair.

4.  **Step 4: Final Formatting and Priority Assignment**
    * **Input:** The flat list of objects from Step 3.
```

```
    * **Process:** For each object, format it according to the `Constraints & Formatting Rules`. Then, calculate a `Priority` number for each entry by ranking them based on two criteria: 1. `"isInOntology"` (`true` before `false`), and 2. `"Retrosynthesis Importance"` (descending). Assign the resulting rank (`1, 2, 3...`) to the `"Priority"` key.
    * **Output:** The final, single JSON object. The list in this JSON does not need to be sorted.

**Constraints & Formatting Rules:**
* The final output **MUST** be a single JSON object. Do not include any text, explanations, or markdown formatting before or after the JSON.
* If no valid disconnections are identified after the full analysis, the output must be a valid JSON object with an empty `disconnections` list (i.e., `{"disconnections": []}`).
* The root key of the object must be `"disconnections"`, containing a list of disconnection objects.
* Each object in the list must contain the following keys:
    * `"disconnection"`: A string representing the complete reaction center **as viewed from the product molecule**. It must list all non-hydrogen atoms **in the product** that are directly involved in the transformation from the reactants. This includes atoms that change their connectivity, atoms whose bonds change order (e.g., a C=C in the reactant becomes a C-C in the product), or atoms that are the site of a stereochemical change. However, for transformations that require adding a new group to the molecule (such as a retrosynthetic protection), you must list the attachment points in the product where the new group is added. The atoms must be separated by spaces.
        * **Example (Bond Cleavage / Deprotection):** `"C:5 N:7"` (These two atoms are bonded in the product but were on separate reactant molecules).
        * **Example (Cycloaddition):** `"c:1 c:2 c:3 c:4 c:5 c:6"` (These six atoms in the product form a new ring that was not present in the reactants).
        * **Example (Functional Group Interconversion - FGI):** `"C:8 C:9"` (Represents a transformation on the bond between these atoms, such as reducing a double bond to a single bond) or `"N:1 O:2 O:3"` (Represents replacing one functional group, like an amine, with its precursor, like a nitro group).
        * **Example (Protection):** `"N:26"` (Represents a transformation at a single or multiple atoms, such as adding a protecting group to an amine nitrogen. For transformations that add a group, this string identifies the single (or multiple) attachment points in the product where the transformation occurs).
        * **Example (Stereochemical Change):** `"C:25"` (This atom in the product has a specific stereochemistry that was set during the reaction).
    * `"Reaction"`: A list representing all reactions of a specific disconnection point. Each individual reaction has:
        * `"forwardReaction"`: A string for the reaction name. If the reaction is from the ontology, use its exact `id`. If you determine that no ontology entry is a good fit and a different reaction is more appropriate (the `OtherReaction` case), you must use your own standard, descriptive name for that reaction (e.g., `"Intramolecular Friedel-Crafts"`).
        * `"isInOntology"`: A boolean (`true` or `false`) indicating if the `"forwardReaction"` name was found in the provided `reaction_ontology` JSON.
        * `"forwardReactionClass"`: The broader reaction class of the `"forwardReaction"` selected from: 'Reduction', 'Acylation', 'Heteroatom Alkylation and Arylation', 'Functional Group Addition', 'Protection', 'C-C Coupling', 'Deprotection', 'Functional Group Interconversion', 'Aromatic Heterocycle Formation', 'Oxidation'. In case of no matching class pick 'Miscellaneous'.
        * `"Retrosynthesis Importance"`: A numerical value from 4 to 1, corresponding to the ranking rationale (4 = Very High, 1 = Lower).
        * `"Priority"`: A sequential integer (`1, 2, 3...`) representing the calculated priority of the disconnection.
        * `"rationale"`: A concise string explaining the strategic value. It must justify the importance level by referencing the strategic goals (a, b, c, d, e), **explicitly state which analysis from Step 1 led to this disconnection** (e.g., 'Convergent disconnection...'), and **comment on any potential chemoselectivity issues, the need for protecting groups, or thermodynamic vs. kinetic control considerations.**
    * **JSON Output Example:**
    {
    "disconnections": [
        {
        "disconnection": "C:1 C:2",
        "reactions": [
            {
            "forwardReaction": "Forward reaction name",
            "isInOntology": true,
            "forwardReactionClass": "Broader reaction class",
            "Retrosynthesis Importance": 4,
            "Priority": 1,
            "rationale": "string"
            },
            // more reactions for the same disconnection point
        ]
        },
        // more disconnection points
    ]
    }

**Retrosynthetic Analysis Framework**
* **Primary Strategic Goals:** Analyze the molecule according to the following framework. Note: You must identify and report reactions on all strategic goal levels. The strategic goals are for the rationale in the final output, not for filtering. Do not omit lesser strategic reactions like protecting group removals.
    * a) **Structural Simplification:** Lead to readily available or simpler starting materials.
    * b) **Reaction Robustness:** Involve robust, high-yielding, and reliable forward reactions.
    * c) **Strategic Construction:** Strategically build the core scaffold or install key functionalities efficiently.
    * d) **Practicality & Efficiency:** Prioritize reactions with good atom economy that avoid notoriously toxic or expensive reagents and are known to be scalable.
    * e) **Stereochemical Control:** For chiral molecules, the plan must address how each stereocenter will be controlled.
* **Ranking Rationale (for assigning Importance value):** Analyze the molecule according to the following framework. Note: You must identify and report reactions from all relevant importance levels. The
```

```
      importance score is for prioritization in the final output, not for filtering. Do not omit lower-
      importance findings like protecting group removals.
    * **Importance 4 (Very High):** Major ring-forming reactions, disconnections that reveal symmetry, or
        those that convergently connect major fragments. Includes powerful skeletal rearrangements that
        build the core.
    * **Importance 3 (High):** Reliable attachment of key functional groups or substituents to an existing
        core. Includes the strategic installation of a key stereocenter via an asymmetric reaction.
    * **Importance 2 (Medium):** Standard functional group interconversions (FGIs) or formation of less
        complex C-C or C-X bonds. Includes less critical rearrangements or stereochemical modifications.
    * **Importance 1 (Lower):** Disconnections of simple, easily accessible fragments or those related to
        reagent synthesis (e.g., protecting groups).
####

**Reaction Ontology:**

<reaction_ontology>

### Molecule for Analysis

**Product SMILES:**

<canonicalized_product>

####

Remember to return all possible reactions. You can identify more than one reaction for a specific position.
```

Listing 1: Position Model Prompt.

A.6.2 TRANSITION MODEL

Note: This prompts is slightly altered for visualization purposes.

```
**Persona:**
You are an expert chemist specializing in synthetic reaction modeling.

**Primary Goal:**
Given a product molecule, a specified reaction center, and a reaction type, your task is to generate all
      chemically reasonable reactant molecules that would form the product. When a reaction name is provided,
      you will model that specific transformation. When it is not, you will suggest and model all plausible
      reactions for the given transformation. You will then validate each option based on practical chemical
      principles. The entire output must be a single, valid JSON object.

**Input Schema:**
* `reaction_center_atoms`: A string identifying the **approximate location** of the transformation, using atom
       mappings. This serves as a guide for the model to identify the precise reaction center.
    * **Example (Bond Cleavage):** `"C:5␣N:7"`
    * **Example (Ring Formation/Cycloaddition):** `"c:1␣c:2␣c:3␣c:4␣c:5␣c:6"`
    * **Example (FGI):** `"C:8␣C:9"`
    * **Example (Protection):** `"N:26"`
    * **Example (Stereochemical Change):** `"C:25"`
* `product_smiles`: The atom-mapped SMILES string of the product molecule.
* `forward_reaction_name` (optional): The name of a specific forward reaction to be modeled.
* `retrosynthesis_reaction_examples` (optional): A list of retrosynthesis reaction SMILES strings to use as a
      blueprint.

**Internal Analysis Pipeline:**
To generate the final JSON object, you will internally execute the following data transformation pipeline.
      This is a strict, one-way sequence from Step 1 to the final output. The steps must be executed exactly
      once in order, without looping back to a previous step. The output of each step serves as the direct
      input for the next.

1.  **Step 1: Determine Reaction(s) to Model**
    * **Input:** The `forward_reaction_name` (optional) and `reaction_center_atoms` from the user.
    * **Process:** If a `forward_reaction_name` is provided, use it as the sole reaction. If not, analyze the
         `reaction_center_atoms` to generate a list of potential `forward_reaction_name`s.
    * **Output (Internal):** A list of reaction names to be modeled.

2.  **Step 2: Refine Reaction Center**
    * **Input:** The list of `forward_reaction_name`s (Step 1), the users `reaction_center_atoms`, and any `
        retrosynthesis_reaction_examples`.
    * **Process:** For each `forward_reaction_name`, use your expert chemical knowledge and the provided
        examples to determine the **precise and complete reaction center**. The users input is a guide for
        the location, but you must refine it by adding or removing atoms to match the true mechanism of the
        reaction.
    * **Output (Internal):** A mapping of each `forward_reaction_name` to its `precise_reaction_center_atoms`
        string.

3.  **Step 3: Extract Atom-Level Reaction Template**
    * **Input:** The list of `forward_reaction_name`s from Step 1, the **precise reaction center** from step
        2, and the user-provided `retrosynthesis_reaction_examples`.
    * **Process:** For each `forward_reaction_name`, analyze its corresponding valid example(s). Your primary
        goal is to extract the **structural pattern** and **JSON format** of the transformation from these
        examples. By analyzing the transformation from the product to the reactant side, extract a formal,
        atom-level retrosynthetic rule (the "template"). If a specific chemical detail in an examples `
        modification_smarts` seems inconsistent with the `forward_reaction_name`, prioritize deriving the
        correct chemical group based on your expert knowledge, while strictly adhering to the JSON structure
         taught by the example. If no valid examples are provided, derive the template from your general
        chemical knowledge.
```

```
    * **Output (Internal):** A mapping of each reaction name to its extracted reaction template. The template
        **must** be a single JSON object following this structure:
        ```json
        // Template Structure: A self-contained rule object
        {
          "precise_reaction_center_atoms": "<space_separated_list_of_atom_maps>",
          "modifications": [
            {
              "target_atom_map": "<map_number_of_atom_to_modify>",
              "modification_smarts": "<SMILES_or_SMARTS_of_the_complete_functional_␣
                   group_on_this_atom_in_the_reactant>"
            }
            // ... one object for each atom that is modified ...
          ]
        }
        ```

    * **Example 1 (Intermolecular Disconnection):** This pattern covers reactions where **one product is
        formed from two** reactant molecules.
        ```json
        {
          "precise_reaction_center_atoms": "C:1␣C:7",
          "modifications": [
            { "target_atom_map": "1", "modification_smarts": "[c:1][X]" },
            { "target_atom_map": "7", "modification_smarts": "[c:7][Y]" }
          ]
        }
        ```

    * **Example 2 (Intramolecular Cyclization):** This pattern covers reactions where a new ring is formed
        within a **single precursor molecule**.
        ```json
        {
          "precise_reaction_center_atoms": "C:1␣C:6",
          "modifications": [
            { "target_atom_map": "1", "modification_smarts": "[C:1]X" },
            { "target_atom_map": "6", "modification_smarts": "[C:6]Y" }
          ]
        }
        ```

    * **Example 3 (Functional Group Interconversion - FGI):** This pattern covers reactions where a functional
         group is transformed into another on a **single molecule**.
        ```json
        {
          "precise_reaction_center_atoms": "C:1␣O:2",
          "modifications": [
            { "target_atom_map": "1", "modification_smarts": "[C:1]=[O:2]" }
          ]
        }
        ```

    * **Example 4 (Multi-Component Reaction - MCR):** This pattern covers reactions where **one product is
        formed from three or more** reactant molecules.
        ```json
        {
          "precise_reaction_center_atoms": "A:1␣B:2␣C:3",
          "modifications": [
            { "target_atom_map": "1", "modification_smarts": "[A]X" },
            { "target_atom_map": "2", "modification_smarts": "[B]Y" },
            { "target_atom_map": "3", "modification_smarts": "[C]Z" }
          ]
        }
        ```

4. **Step 4: Generate Precursor Molecule(s)**
    * **Input:** The `product_smiles` and `precise_reaction_center_atoms`.
    * **Process:** Based on the number of fragments implied by the transformation type (e.g., two for an
        intermolecular disconnection, one for an FGI, three for a 3-component MCR), generate the
        corresponding core precursor molecule(s). This is done by cleaving the necessary bonds in the
        product or, for 1-to-1 transformations, identifying the single precursor scaffold.
    * **Output (Internal):** The distinct molecular fragment(s) with atom mapping preserved.

5. **Step 5: Apply Reaction Template to Generate Reactant Permutations**
    * **Input:** The precursor(s) (Step 4) and the reaction templates (Step 3).
    * **Process:** For each reactions template, apply the extracted retrosynthetic template to the precursor(s
). The `precise_reaction_center_atoms` provided by the user defines the **locality** of the
        transformation. You must use your chemical expertise to apply the template correctly to the atoms **
        in and around this specified location**, ensuring the final transformation is chemically consistent
        with the templates logic. This process must include generating **all possible permutations** of the
        reactive groups. This directive must be interpreted with absolute completeness in two ways:
        1. **Fragment-Role Permutations:** For a disconnection into multiple fragments with distinct reactive
             groups, you must generate reactant sets for **all** possible assignments of those groups to the
             fragments.
        2. **Intra-Group Class Permutations:** If a generated reactive group belongs to a general chemical
             class (e.g., an "organohalide," "leaving␣group," or "protecting␣group"), you are required to
             generate an exhaustive list of separate options for **all chemically distinct members of that
             class known to be compatible with the reaction.**
        The model is **explicitly forbidden** from filtering this list based on commonality, synthetic
             efficiency, or perceived viability. If a variant is chemically possible, it must be included in
             the output.
```

```
    * **Output (Internal):** A list of all potential reactant options generated from this exhaustive process,
        each associated with a `forward_reaction_name`. No chemically possible permutations may be omitted.
        Please dont provide reagents as reactants.

6.  **Step 6: Validate and Justify Each Option**
    * **Input:** The list of potential reactant options from Step 5.
    * **Process:** For each generated option, perform a rigorous chemical validation.
        * A) **Stability:** Are the proposed reactants chemically stable?
        * B) **Chemoselectivity:** Would the reaction be selective? Are there other functional groups that
            would interfere?
        * C) **Stereochemical Consistency:** Is the transformation stereochemically sound? Does it correctly
            account for the creation or modification of stereocenters in the product?
        * D) **Plausibility:** Is the reaction electronically and sterically plausible for this specific pair?
    * **Output (Internal):** The same list of options, but now each object contains an `is_valid` boolean and
        a detailed `reasoning` string that explicitly addresses these validation points.

### **Step 7: Final Formatting and Grouping**
* **Input:** The validated and justified flat list of *real chemical options* from Step 6.
* **Process:**
    1.  **Group Options:** Begin by grouping the list of validated options by their `forward_reaction_name`.
    2.  **Extract Wildcard Reaction Class** Looking at the validated options and their reaction names, you
        must deduct a general reaction class template if possible using the `<CLASS:..>` tag. It signals
        that a member of this chemical class (e.g. `<CLASS:AmineProtectingGroup>`) should be used instead of
         an explicit molecular structure.
    3.  **Generate General Template Entry (if applicable):** For each extracted general reaction class
        template, you **should** create one additional, special permutation object derived from the two
        provided general reaction classes. This object serves as the general, machine-readable
        representation for the entire transformation class and should be placed at the **beginning** of the
        `reactant_permutations` list. The two possible options for this general reaction class template are:
        * For a **Defined Chemical Class** (e.g., `<CLASS:Halogen>`), where the reactants share a specific
            generalizable atoms across all precursor molecule(s) from Step 6, introduce the a SMARTS pattern
             (e.g., `[A,B,C]`) as a replacement for these generalizable atoms. If possible, create a joined
            template covering generalizable atoms on all possible reactants instead of creating multiple
            templates.
        * For a **Wildcard Addition Class** (e.g., `<CLASS:ProtectingGroup>`), where the specific reagent
            added in the retrosynthetic step is a strategic choice from a broad and variable unknown set,
            the added group is represented by a generic wildcard atom (`[*]`). This string is generated by
            taking the appropriate precursor molecule(s) from Step 6 and creating a new bond between the
            wildcard atom (`[*]`) and the product that generalizes the explicit reactant options.
        * This special permutation object must have the following structure:
            * `reactants`: A list containing the single, atom-mapped SMILES string with the general
                representation.
            * `is_valid`: `true`.
            * `is_template`: `true`. Indicating that this result is a wildcard template.
            * `reasoning`: A string that explicitly identifies this as the general template and names the
                chemical class in the format `<Class:XYZ>`.
    4.  **Assemble Final List:** For each unique reaction, create a single object containing the `
        forward_reaction_name` and its final `reactant_permutations` list. This list will now contain the
        general template entry at the top (if applicable), followed by all the validated, specific examples
        from Step 6.
    5.  **Finalize and Clean:** Assemble these grouped objects into the final `reaction_analysis` list
        according to the `Output Schema`. Keep the original atom mapping of the product where possible and
        do not introduce new atom maps on the reactant side, but use unmapped atoms.
* **Output:** The final, single JSON object.

**Output Schema     Strict JSON Only:**
'''json
{
  "product": "<SMILES>",
  "reaction_analysis": [
    {
      "forward_reaction_name": "Name␣of␣Reaction␣1␣(e.g.,␣Suzuki-Miyaura␣coupling)",
      "reactant_permutations": [
        {
          "reactants": ["<SMILES_1A>", "<SMILES_1B>"],
          "is_valid": true,
          "is_template": false,
          "reasoning": "This␣permutation␣is␣valid.␣The␣reactants␣are␣stable␣and␣the␣reaction␣is␣chemoselective
              ."
        },
        {
          "reactants": ["<SMILES_2A>", "<SMILES_2B>"],
          "is_valid": false,
          "is_template": false,
          "reasoning": "This␣permutation␣is␣invalid␣due␣to␣severe␣steric␣hindrance␣at␣the␣reaction␣site."
        }
      ]
    }
    // ... one object for each unique reaction suggested in Step 1 ...
  ]
}

** Input **

"reaction_center_atoms": <REACTION_POSITION>
"forward_reaction_name": <REACTION_NAME>
"product_smiles": <PRODUCT_SMILES>
"retrosynthesis_reaction_examples": <TRAIN_REACTION_EXAMPLES>
```

Listing 2: Transition Model Prompt.

```
Task:
Given a product molecule, a reaction center, and an optional reaction name, your task is to generate all
      chemically reasonable reactant molecules that would form the product. The entire output must be a single
      , valid JSON object following the specified schema.

Instructions:

    Identify the reaction(s) to model based on the inputs.

    For each reaction, determine the retrosynthetic disconnection.

    Generate all possible reactant permutations, including variations for chemical classes (e.g., all halogens
          for an organohalide). Do not filter out any chemically possible options.

    For each permutation, validate its chemical feasibility (stability, selectivity, etc.) and provide a brief
          justification.

    Group the results by forward_reaction_name in the final JSON output.

Input Schema:

    reaction_center_atoms: A string identifying the approximate location of the transformation, using atom
         mappings.

        Example (Bond Cleavage): "C:5 N:7"

        Example (Ring Formation/Cycloaddition): "c:1 c:2 c:3 c:4 c:5 c:6"

        Example (FGI): "C:8 C:9"

        Example (Protection): "N:26"

        Example (Stereochemical Change): "C:25"

    product_smiles: The atom-mapped SMILES string of the product molecule.

    forward_reaction_name (optional): The name of a specific forward reaction to be modeled.

    retrosynthesis_reaction_examples (optional): A list of retrosynthesis reaction SMILES strings to use as a
         blueprint.

Output Schema     Strict JSON Only:

{
  "product": "<SMILES>",
  "reaction_analysis": [
    {
      "forward_reaction_name": "Name of Reaction 1 (e.g., Suzuki-Miyaura coupling)",
      "reactant_permutations": [
        {
          "reactants": ["<SMILES_1A>", "<SMILES_1B>"],
          "is_valid": true,
          "is_template": false,
          "reasoning": "This permutation is valid. The reactants are stable and the reaction is chemoselective
               ."
        },
        {
          "reactants": ["<SMILES_2A>", "<SMILES_2B>"],
          "is_valid": false,
          "is_template": false,
          "reasoning": "This permutation is invalid due to severe steric hindrance at the reaction site."
        }
      ]
    }
    // ... one object for each unique reaction suggested ...
  ]
}



** Input **

"reaction_center_atoms": <REACTION_POSITION>
"forward_reaction_name": <REACTION_NAME>
"product_smiles": <PRODUCT_SMILES>
"retrosynthesis_reaction_examples": <TRAIN_REACTION_EXAMPLES>
```

Listing 3: Ablation Study Short Transition Model Prompt.