ACL Supplementary Material

1 Data Construction Details

The data generation script and associated prompt template has been made publicly available in *CoT-Gen.py* and *User-Prompt.txt*, respectively. And a subset of 100 randomly selected examples is provided in *Open-Source.jsonl* for demonstration purposes.

1.1 Forward Prediction

This is the prompt template used in Section **Stage 1: Preliminary Construction** to generate forward prediction CoT data. The prompt first provides explicit high-level instructions, guiding the model to assume the role of an expert chemist and to articulate the reaction reasoning process in a step-by-step manner.

Prompt Template

You are an expert AI-chemist capable of explaining the reasoning process for chemical reactions step by step.

You will be given a standardized chemical reaction equation, including reactants, reagents, and products. Your task is to construct a reasoning dataset based on this equation.

For each reasoning step, provide a title describing what you did in that step and its content.

Requirements: Use 5-8 reasoning steps. Verify the given answer through different methods. Consider potential alternative answers and explain why they are rejected. Contemplate the possibility of error: If your reasoning is wrong, identify where the error could occur. Fully test all other possibilities where you might be mistaken. When you say you are rechecking, genuinely recheck using another method instead of merely stating that you are rechecking. Format your response in JSON. Each reasoning step should include the following keys:

"title": A description of the current reasoning step.

"content": A detailed explanation of this step.

"next action": Either "continue" or "final answer".

Subsequently, a concrete reaction example is included to demonstrate the expected reasoning structure and output format, enabling the LLM to imitate both the logical progression and the structured JSON representation during data generation. The textual descriptions corresponding to the key reaction mechanisms are highlighted in bold.

Reaction Reasoning Example

Reaction inputs:

CN(C)c1ccc(-c2cnc3ccc(O[Si](C)(C)C(C)(C)C)cc3n2)cc1.[F-]

Reaction output:

CN(C)c1ccc(-c2cnc3ccc([OH])cc3n2)cc1

"title": Analyzing the given reactants and reagents

"content": The reactants and reagents include

CN(C)c1ccc(-c2cnc3ccc(O[Si](C)(C)C(C)(C)C)cc3n2)cc1, a molecule with a central aromatic ring substituted with a pyridine moiety, and [F-], a fluoride ion.

"next-action": "continue"

"title": Hypothesizing the possible reaction mechanism

"content": The reaction may involve the following reaction centers: The silicon atom in the silyl ether group

and its attached oxygen atom; The incoming fluoride ion interacts with the silicon. The reaction describes the transformation of a silyl ether to a hydroxyl group. **This process typically involves a deprotection reaction where the silyl protecting group is removed, exposing the original alcohol**. Here, the fluorine ion ([F-]) acts as a nucleophile in the mechanism, which is common for desilylation reactions.

"next-action": continue

"title": Predicting possible reactions

"content": Here are the possible mechanistic steps of this reaction. Nucleophilic attack: The fluoride ion attacks the silicon atom of the silyl ether. Silicon has a high affinity for fluorine due to its ability to form strong Si-F bonds. **Cleavage of Si-O bond:** As the Si-F bond forms, the Si-O bond is weakened and subsequently cleaved. This releases the Si group (usually as a volatile SiF compound), and the hydroxyl group is restored. **Reformation of hydroxyl group:** With the removal of the silyloxy group, a hydrogen from the solvent or environment can bond to the oxygen, completing the transformation to the hydroxyl group.

"next-action": continue

"title": Validating the reaction pathway

"content": Fluoride ion (F-) acts as a nucleophile and attacks the silicon atom in the TBDMS group. Upon attack by fluoride, the silicon center forms a pentavalent transition state, which is relatively stable because silicon can accommodate five bonds due to its size and electronic structure. This intermediate is often depicted with a negatively charged fluoride attached to silicon, briefly forming a complex where the oxygen-silicon bond is weakened. As the intermediate collapses, **the Si-O bond between the silyl group and the oxygen atom is cleaved**, breaking the silyl ether linkage and releasing a free alcohol (-OH). After the Si-O bond has been cleaved, **the molecule is left with a free hydroxyl group (-OH) in place of the silyl ether.** The Si-F bond formation results in the formation of tert-butyldimethylsilyl fluoride (TBSF) as a byproduct. TBSF is relatively stable and does not interfere with the reaction further.

"next-action": continue

"title": Considering alternative possibilities

"content": The first possible reaction is that the fluoride ion could attack the electron-deficient aromatic ring near the nitrogen atoms, potentially displacing a substituent in a nucleophilic aromatic substitution. However, Nucleophilic aromatic substitution (SNAr) generally requires strong electron-withdrawing groups ortho or para to the leaving group, which are not present in this molecule. Fluoride is more likely to target the silicon atom in the silyl group due to silicon's affinity for fluoride and the relatively low activation energy required for Si-O bond cleavage. The second possible reaction is that the nucleophilic fluoride could theoretically attack the methyl groups attached to the tertiary nitrogen, leading to N-dealkylation and removing one or both methyl groups. However, N-dealkylation would typically require a more electrophilic carbon (such as a benzylic or allylic carbon) or harsher conditions, such as oxidation. In contrast, the silicon-fluoride interaction is very strong, making the silyl ether bond much more susceptible to fluoride attack than the N-C bonds in the tertiary amine. Fluoride ions could, under special conditions, participate in deprotonating an aromatic C-H bond or causing other activation of the aromatic ring. However, Aromatic C-H activation is highly unlikely under these conditions and would generally require a transition-metal catalyst and more stringent conditions. The Si-O bond in the silyl ether is far more reactive toward fluoride under mild conditions. The last possible reaction could be a fluoride-mediated deprotection of a tert-butyldimethylsilyl (TBDMS) group, restoring the free hydroxyl (-OH) functionality in the aromatic compound. The high affinity between silicon and fluoride drives the reaction forward. The formation of the strong Si-F bond makes the reaction favorable and specific for silyl deprotection. The fluoride ion selectively attacks the silicon center due to its high electrophilicity, leaving other functional groups (like the tertiary amine and aromatic rings) untouched.

"next-action": continue

"title": Confirming the reasoning process

"content": After reasoning and validation,
CN(C)c1ccc(-c2cnc3ccc([OH])cc3n2)cc1 is confirmed to be the final product.

"next-action": continue

"title": Final reasoning result

"content": Based on the above analysis, the reaction product is
CN(C)c1ccc(-c2cnc3ccc([OH])cc3n2)cc1. Instead of the silyl ether group, there is a hydroxyl (OH) group.

"next-action": final-answer

In addition, a visual representation of the reaction with annotated reactive sites, together with the corresponding SMARTS reaction template, is provided here for illustration purposes (Fig. 1) to complement the textual reasoning and confirm the correctness of the key reaction mechanisms described therein.

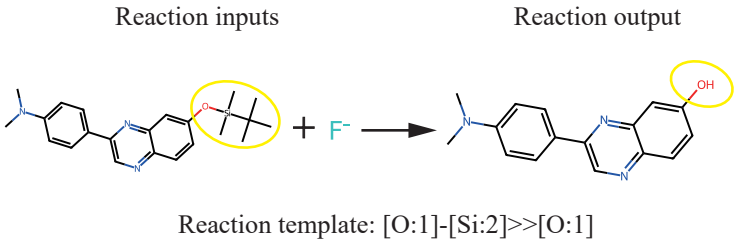

Figure 1: Visual representation of the reaction with annotated reactive sites, together with the corresponding SMARTS reaction template. This reaction involves the cleavage of the Si–O bond, followed by the reformation of the hydroxyl group.

1.2 Retrosynthesis Prediction

This is the prompt template to generate retrosynthesis prediction reasoning data. Here we provide a core thinking step template rather than specific examples.

Prompt Template

You are an expert in organic chemistry.

Given the complete standard reaction equation, including reaction inputs and products.

You need to reason through the process. Given the target product, pretend not to know the reaction inputs, and deduce the required reaction inputs (which may include reactants, reagents, solvents, etc.). Do not reveal the standard answer prematurely. The final result must match the standard answer.

The reasoning content must strictly follow the format below, with each reasoning step title placed in <>.

<think>

1. <Analyze the target product>

2. <Analyze bond-disconnection strategy>

3. <Predict reaction inputs>

4. <Validate the synthetic route>

5. <Confirm reaction inputs> (reaction inputs given in the form <chem>SMILES</chem>)

</think>

6. <Give final result>

2 Method and Experiment Details

Due to page constraints, we provide here a detailed supplement to the section **Method** and **Experiment** in the main manuscript.

2.1 About RAG

Here, we expand on the details of the RAG methodology and experimental settings.

2.1.1 Parameter Selection

In our experiments about reaction type classifier, we first observed that the classifier exhibited differing performance on the forward prediction and retrosynthesis prediction tasks. For forward prediction (the

reaction inputs are given to predict the reaction type) the Top-1 to Top-3 accuracy were 94.4%, 97.2% and 99.0%, respectively. In contrast, for retrosynthesis prediction (the target product is given to predict the reaction type) the corresponding Top-1 to Top-3 accuracy were 69.5%, 75.2% and 83.9%. The lower accuracy observed for the retrosynthesis prediction task aligns with chemical intuition: It is challenging to determine the reaction type of a synthetic pathway based solely on the final product molecule.

Given these results, we adopt different retrieval strategies for the two tasks. For forward prediction, we retrieve similar reaction cases only from the Top-1 predicted type, selecting five cases. For retrosynthesis prediction, we retrieve cases from the Top-3 predicted types, selecting two cases from each type, resulting in a total of six retrieved examples.

2.1.2 New Baseline for Comparison

Our subsequent algorithm comparisons are based on CoT data after the second-round of data scaling, using the method detailed in section **Scaling High-Quality Reasoning Data** (with its first-round performance gains shown in Table 2).

2.2 About Reward Adjustment

2.2.1 Reward Modes and Training Performance

This is the supplementary to Figure 2 in the main text. We define three different reward modes for GRPO and plot the corresponding reward curves, with Figure 2A and Figure 2B showing the reward curves for forward reaction prediction and retrosynthesis prediction tasks, respectively.

- **Mode-1** employs the basic reward configuration, which comprises **Format Reward**, **Validity Reward**, **Length Reward**, and **Accuracy Reward**, with their weights all fixed at 1.0, respectively. Its validation performance is evaluated using the Exact Match metric and depicted by the blue curve.
- **Mode-2** extends this setup by introducing an additional **Feasibility Reward**, whose weight is fixed at 0.5 throughout training while all other rewards remain unchanged. Its validation performance is evaluated using the Feasibility metric and depicted by the yellow curve.
- **Mode-3** further modifies Mode-2 by applying a dynamic weighting strategy to **Feasibility Reward** during training. Its validation performance is also evaluated using the Feasibility metric and depicted by the green curve.

We find that the adopted method performs well in both forward prediction and retrosynthesis prediction tasks. Although the performance curves show little distinction between Mode-2 and Mode-3, Table 5 in the main text demonstrates that the dynamic reward weighting strategy yields a notable performance improvement.

2.2.2 Multi-Stage Tuning with Fixed Rewards

We experimented with numerous hyperparameter combinations and gradually refined them based on observed experimental behavior. Among all settings, the choice of step K proved most critical. For reaction prediction tasks, we found that the LLM tends to explore the reaction space relatively early during GRPO training, and continuing to emphasize the **Feasibility Reward** thereafter can lead to overfitting to this specific reward and a substantial decline in Exact Match (%) performance. Consequently, we set K toward the beginning of training and found that a value of 300 yields a improvement in performance (shown in Table 4, compared with result using RAG in Table 3).