# MiST: Understanding the Role of Mid-Stage Scientific Training in Developing Chemical Reasoning Models

**Andres M Bran[*1] Tong Xie[*2] Shai Pranesh[1] Jeffrey Meng[1] Xuan Vu Nguyen[1] Jeremy Goumaz[1] David Ming Segura[1] Ruizhi Xu[2] Dongzhan Zhou[2] Wenjie Zhang[2] Bram Hoex[2] Philippe Schwaller[1]**

*[*]Equal contribution [1]EPFL, Switzerland [2]UNSW Sydney, Australia*

*Correspondence to: Philippe Schwaller philippe.schwaller@epfl.ch; Tong Xie tong.xie@unsw.edu.au*

## 1. Introduction

Large Language Models can develop reasoning capabilities through online fine-tuning with rule-based rewards. However, recent studies reveal a critical constraint: reinforcement learning (RL) succeeds only when the base model already assigns non-negligible probability to correct answers—a property we term "latent solvability." This work investigates the emergence of chemical reasoning capabilities and what these prerequisites mean for chemistry.

We identify two necessary conditions for RL-based chemical reasoning: (1) Symbolic competence—models must read and generate syntactically valid chemical strings like SMILES, IUPAC names, or CIF files; and (2) Latent chemical knowledge—answers must exist in the long-tail of the model's prior distribution, so that RL training can exploit them.

## 2. Mid-Stage Scientific Training (MiST)

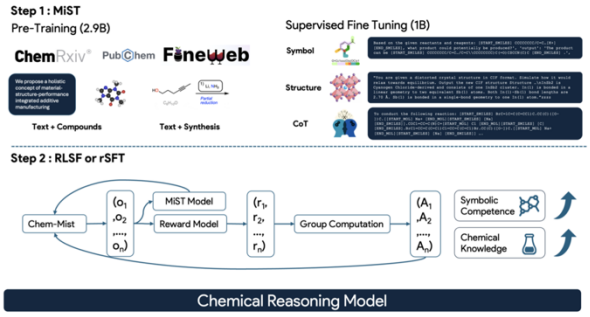


Figure 1 Multi-stage pipeline for training a chemical-reasoning language model.

Step1 (MiST, 3.9 B tokens) Continued Pretraining exposes a general-purpose base model to a chemistry-centric corpus that interleaves plain text with compound \& synthesis information. A subsequent 1 B-token supervised fine-tuning phase teaches three formats: (i) symbol-level molecular or material understanding, (ii) structure-aware question \& answers, and (iii) chemical chain-of-thought (CoT). In Step2 the MiST backbone is further specialized with either RLSF (reinforcement learning from scientific feedback) or rSFT (reasoning-style supervised fine-tuning). A pool of candidate answers generated by the MiST model is scored by a task-specific reward model ;a group-computation module aggregates these signals to update the policy, iteratively refining the model into a \emph{Chemical Reasoning Model}.

We propose Mid-Stage Scientific Training (MiST): a set of mid-stage training techniques to satisfy these prerequisites. MiST includes: (i) data-mixing with SMILES/CIF-aware pre-processing to maintain chemical notation integrity, (ii) continued pre-training on 2.9B tokens of scientific literature, and (iii) supervised fine-tuning on 1B tokens of curated chemistry tasks.

Our preprocessing pipeline leverages Nougat and GROBID for PDF-to-text conversion, with superior performance in transforming complex structures such as tables, formulae, and bibliographic references into LaTeX-formatted text while preserving chemical compounds with their corresponding SMILES representations.

### *2.1 Related Work*

Prior work on chemical LLMs has focused on either molecular representations [1] or reasoning capabilities [2] in isolation. Our approach bridges this gap by identifying that both symbolic competence and latent knowledge are prerequisites for successful RL-based reasoning. Unlike approaches using SELFIES for guaranteed validity, our diagnostic benchmarks demonstrate the importance of learnable validity constraints.

## 3. Results

MiST raises the latent-solvability score on 3B and 7B parameter models by up to 1.8×. When combined with RL, we observe dramatic improvements: top-1 accuracy increases from 10.9% to 63.9% on organic reaction naming, and from 40.6% to 67.4% on inorganic material generation (Conditional Material Generation task).

The accuracy reward evolution during RL training shows the base Qwen2.5-3B model

plateaus early at a reward below -0.5, indicating frequent generation of syntactically valid SMILES but incorrect products. In contrast, both fine-tuned variants (Qwen2.5-3B+SFT and Qwen2.5-3B+MiST+SFT) maintain accuracy rewards above -0.5 throughout training.

Our ablations confirm that removing any single prerequisite collapses RL gains. Pretrained models like Qwen2.5-3B lack symbolic abilities needed for SMILES understanding, but this is overcome with MiST. The CMG task for inorganic crystals exhibits distinct behavior: while RL leads to decreased symbolic competence (model shifts toward CIF-derived sequences), it yields marked increases in chemical knowledge specific to inorganic materials.

| | Metrics | | Reasoning tasks | | | |
|---|---|---|---|---|---|---|
| Model | SCS ↑ | CCS ↑ | I2S ↑ | RxP ↑ | RxN ↑ | CMG ↑ |
| **Qwen-2.5 3B** | 0.955 | 0.352 | 0.0 | 0.0 | 10.33 (10.9) | 40.6 (44.4) |
| +CP | 1.561 | 0.404 | 1.0 | 0.0 | 11.1 (10.3) | 40.1 (47.8) |
| +SFT | 1.650 | 0.707 | **52.7** | 5.2 (13.6) | 15.1 (17.5) | 38.9 (46.6) |
| +RL(I2S) | 1.535 | 0.695 | 52.0 | 3.2 (12.2) | 13.3 (17.2) | 42.1 (58.3) |
| +RL(RxP) | 1.880 | 0.782 | 49.9 | **6.6 (17.4)** | 14.5 (16.9) | 40.3 (49.2) |
| +RL(RxN) | 1.650 | 0.698 | 48.9 | 5.8 (9.0) | **28.5 (46.8)** | 43.6 (51.8) |
| +RL(CMG) | 0.119 | 1.737 | 50.4 | 0.0 (7.8) | 13.1 (18.2) | **64.9 (70.5)** |
| **Qwen-2.5 7B** | 0.97 | 0.406 | 0.0 | 0.0 (0.2) | 10.7 (12.1) | 40.8 (43.5) |
| +CP | 1.67 | 0.445 | 0.2 | 0.8 (0.4) | 14.8 (14.7) | 45.6 (45.8) |
| +SFT | 1.74 | 0.775 | **65.7** | 13.2 (25.2) | 13.8 (30.1) | 38.2 (57) |
| +RL (I2S) | 1.67 | 0.766 | 65.2 | 12.6 (25.2) | 22.7 (31.4) | 38.5 (52.6) |
| +RL (RxP) | 1.71 | 0.770 | 65.2 | **15.6 (29.8)** | 11.7 (31.2) | 39.6 (52.1) |
| +RL (RxN) | 1.73 | 0.731 | 61.7 | 13.2 (12.6) | **26.4 (63.9)** | 23.6 (52.1) |
| +RL (CMG) | 0.885 | 1.869 | 65.72 | 14.6 (15.8) | 13.8 (29.6) | **67.4 (73.1)** |
| **Ablations (7B)** | | | | | | |
| only RL (RxP) | 1.03 | 0.408 | 0.0 | 0.0 (0.0) | 9.97 (12.1) | 0.0 (47.3) |
| **Baselines** | | | | | | |
| ChemLLM-7B | 1.18 | — | 0.5 | 2.04 (—) | 18.7 (—) | 20 (—) |

Figure 2 Effect of MiST and each post-training stage on downstream reasoning tasks. SCS = symbolic-competence score, CCS = chemical-competence score; both are unitless effect-size measures ranging from 0 (no separation) to 2 (near-perfect separation); higher is better, see Section 3.1. I2S = IUPAC→SMILES translation, RxP = forward reaction prediction, RxN = reaction-naming, CMG = conditional material generation. For the four downstream tasks we report top-1 accuracy. The value outside the parentheses is obtained with a "direct answer" (no-chain-of-thought) prompt. Values inside parentheses are the accuracy when "reasoning" (chain-of-thought) prompting is induced.

## 4. Conclusions

We demonstrate that targeted mid-stage pre-training unlocks chemical reasoning in smaller models. Our findings provide a framework to inform how reasoning-oriented RL methods will perform when applied to complex scientific domains. Similar diagnostic metrics can be devised for biological sequential data, mathematical notation, and other symbolic scientific representations. Future work should address limitations including syntactic-focused rewards and extension to broader chemical representations.

## Acknowledgments

This work was supported by EPFL and UNSW Sydney. We thank the open-source community for providing the Qwen model family and associated tooling.

## References

[1] Weininger, D. SMILES, a chemical language and information system. J. Chem. Inf. Comput. Sci., 28:31–36, 1988.

[2] Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. NeurIPS, 35:24824–24837, 2022.

[3] Blecher, L. et al. Nougat: Neural Optical Understanding for Academic Documents. arXiv preprint arXiv:2308.13418, 2023.

[4] Coley, C.W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem. Sci., 10(2):370–377, 2019.

[5] Kim, S. et al. PubChem 2025 update. Nucleic Acids Research, 53(D1):D588–D597, 2025.