
Supplementary Information for MiST: Understanding the role of Mid-stage Scientific Training in developing chemical reasoning models

1 Chemical Tasks

This section compiles all of the chemical tasks we have implemented for training models with RLVR, including details of the dataset compilation, rewards used, along with additional experimental results not shown in the main manuscript. The details for each task are specified with a table containing a summary of the rewards used, short explanations of what the input and expected outputs are, and finally, some sample outputs from the trained models. Following the table, we report additional experimental results, implementation details, and other task-specific details.

1.1 Reaction Prediction (RxP)

<i>Task: Reaction prediction</i>	
Dataset Source	USPTO-480K ²
Input	Reason and predict the correct product in SMILES notation from the following reaction: <chem>CCN(CC)CC.C0c1ccc(O)cc1.C1C(C1)Cl.C1CC1.O.O=C(C1)c1ccccc1.</chem>
Target Molecule	<chem>C0c1ccc(OC(=O)c2ccccc2)cc1</chem>
Output	<think> The reaction involves the acylation of a phenolic hydroxyl group using benzoyl chloride <chem>O=C(C1)c1ccccc1</chem> in the presence of triethylamine <chem>CCN(CC)CC</chem> as a base. Here’s the step-by-step mechanism: ... 1. Reactants Identification**: ... 2. Mechanism**: ... 3. Product**: ... Final Answer**: The product SMILES is <chem>C0c1ccc(OC(=O)c2ccccc2)cc1</chem> representing the acetylation of the phenolic oxygen [O] in 4-methoxyphenol <chem>C0c1ccc(O)cc1</chem> by benzoyl chloride <chem>O=C(C1)c1ccccc1</chem>, facilitated by triethylamine <chem>CCN(CC)CC</chem>. </think> <answer> <chem>C0c1ccc(OC(=O)c2ccccc2)cc1</chem> </answer>
Answer	<chem>C0c1ccc(OC(=O)c2ccccc2)cc1</chem> ✓

Table 1: Example of reaction prediction task.

The reaction prediction task requires the model to reason and predict the correct product molecule given a list of reactants and reagents (Table 1). Solving this task usually requires expert chemists to think about the reactivity of the reactants involved, and propose and evaluate different reaction mechanism hypotheses. These serve as arguments and causal explanations that support the decisions.

The dataset for the RLVR training of this task was derived from the USPTO-480K² after removing the samples used in the SFT phase. 50K reactions were randomly chosen for the training set, and 500 reactions for the test set.

Given a model output o , from which a final answer a can be extracted, the reward function is the sum of format correctness ($R_{\text{format}} : o \mapsto [-1, 1]$, see Appendix 3) and accuracy of the predicted product ($R_{\text{acc}} : a \mapsto \{-1, -0.5, 1\}$). The accuracy reward is determined by an exact match check against the ground truth:

$$R_{\text{acc}}(a) = \begin{cases} -1, & \text{if Ans cannot be captured from Output or is not a valid SMILES.} \\ -0.5, & \text{if Ans refers to a molecule different than the ground truth.} \\ +1, & \text{if Ans corresponds to the ground truth molecule.} \end{cases}$$

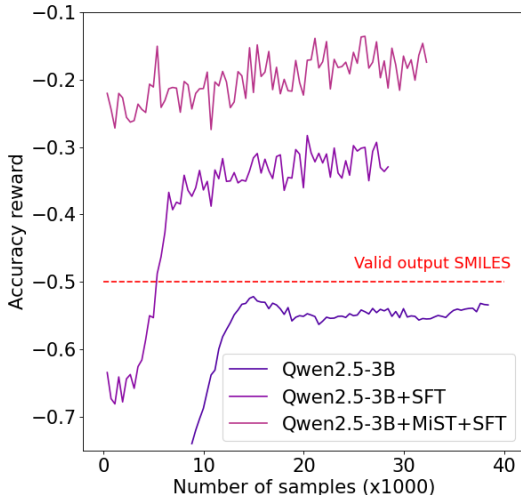


Figure 1: Accuracy reward evolution.

Figure 1 illustrates the evolution of the accuracy reward throughout training. The base Qwen2.5-3B model plateaus early at a reward below the -0.5 threshold, indicating that while it frequently generates syntactically valid SMILES strings, it fails to predict the correct product molecules. In contrast, both fine-tuned variants (Qwen2.5-3B+SFT and Qwen2.5-3B+MiST+SFT) maintain accuracy rewards above -0.5 in the majority of the training process. The SFT-only model shows a sharp increase in reward during the first 8,000 training samples, reflecting rapid format alignment. This is followed by a more gradual rise, suggesting improved chemical reasoning over time. Notably, the model pretrained with MiST before SFT follows a similar trajectory but consistently achieves a higher reward ceiling. This highlights the effectiveness of MiST and SFT in enhancing the model’s learning capability during RLVR.

1.2 IUPAC to SMILES (I2S)

Table 2: Example of IUPAC→SMILES conversion.

<i>Task: IUPAC to SMILES Conversion</i>	
Dataset Source	PubChem (CID 2,000,001 to 3,000,000) ⁸
Input	Question: You are an expert in Cheminformatics, who is very familiar with Simplified Molecular Input Line Entry System (SMILES) notation, and here’s a task for you: Given a molecule with the IUPAC name as below, please provide the corresponding SMILES notation. Here is the IUPAC name: "4-(azepan-1-ylsulfonyl) -N-(pyridin-3-ylmethylideneamino) benzamide".
Target Molecule	<chem>O=C(NN=CC1=CC=CN=C1)C1=CC=C(S(=O)(=O)N2CCCCC2)C=C1</chem>
Output	<think> The SMILES representation of 4-(azepan-1-ylsulfonyl) -N-(pyridin-3-ylmethylideneamino) benzamide is <chem>O=C(NN=CC1=CC=CN=C1)C1=CC=C(S(=O)(=O)N2CCCCC2)C=C1</chem> ...</think>
Answer	<answer> <chem>O=C(NN=CC1=CC=CN=C1)C1=CC=C(S(=O)(=O)N2CCCCC2)C=C1</chem> </answer> ✓

The IUPAC-to-SMILES conversion task presented requires the model to translate a systematic chemical name into its corresponding SMILES string notation. We source our examples from the USPTO-480K dataset and prompt the model with a clear generative instruction “Given this IUPAC name, please provide the corresponding SMILES notation.” The ground truth is established by the Tanimoto similarity score. The model’s chain-of-thought reasoning is captured inside <think>...</think> tags and the final SMILES inside <answer>...</answer>. This setup tests the model’s ability to parse complex nomenclature and generate SMILES structural representations.

For GRPO reward pipeline, we use accuracy reward function that only considers the final ‘<answer>’ SMILES s_{ans} . After converting both the reference s^* and prediction s_{ans} to Morgan fingerprints, it computes the raw Tanimoto similarity τ and then applies a piecewise linear scaling:

$$r_{\text{tf}}(\tau) = \begin{cases} +1.0, & \tau = 1.0, \\ \tau - 0.3, & 0.3 \leq \tau < 1.0, \\ -0.5, & \tau < 0.3. \end{cases}$$

A perfect match yields $r_{\text{tf}} = 1.0$, moderate similarity is rewarded proportionally above -0.5 , and very poor matches incur a negative penalty.

1.3 Reaction Naming

Task: Reaction Naming (RxN)

Dataset Source	USPTO Reaction 1M; ¹
Input	What is the name of this chemical reaction? The reagents are <chem>CC(C)(C)[Si](C)(C)Cl</chem> and <chem>Nc1ccc(-c2cccc(O)c2)c2c1C(=O)NC2</chem> and the product <chem>CC(C)(C)[Si](C)(C)Oc1ccc(-c2ccc(N)c3c2CNC3=O)c1</chem>. Choose ONLY from the following options and write your response choice inside <code><answer>...</answer></code>: "Acylation", "Aromatic Heterocycle Formation", "C-C Coupling", "Deprotection", "Functional Group Addition", "Functional Group Interconversion", "Heteroatom Alkylation and Arylation", "Miscellaneous", "Protection", "Reduction". Do not provide final answer different than what is provided in this list.
Target Molecular Reaction Name	Protection
Output	<code><think></code> First, let's break down the reagents. The first reagent is <chem>CC(C)(C)[Si](C)(C)Cl</chem> which is tert-butyltrimethylsilyl chloride ...
Answer	<code><answer></code> Protection <code></answer></code>

Table 3: Reaction Naming experiment description.

The reaction naming task is a classic example of a structured classification problem in cheminformatics, where the goal is to categorize the nature of a reaction given reactants, conditions and products. This approach aims to test the ability of the LLM to conduct chemical reasoning and instruction following for discrete level answering. In addition to that, this setup also tests the model's ability to interpret chemical structures from linear notation and enables us to reveal how chain-of-thought guidance and prompt design impact classification accuracy. To stimulate reasoning, the model is tasked to output his thinking process inside `<think>...</think>` tags before emitting the final choice in `<answer>...</answer>` tags. The ground-truth class labels are evenly drawn from ten commonly found reaction type in chemistry: "Acylation", "Aromatic Heterocycle Formation", "C-C", "Coupling", "Deprotection", "Functional Group Addition", "Functional Group Interconversion", "Heteroatom", "Alkylation and Arylation", "Miscellaneous", "Protection" and "Reduction" derived from curated USPTO reactions dataset.

Reward Functions:

- **Continuous Format Reward:**
 - This reward is described in Section 3.2.1 in the Algorithm 3.
- **Accuracy Reward:**
 - 0 if no answer is given
 - 0.1 if a single answer is given (but wrong)
 - 1 if the answer is entirely correct
 - -0.2 penalty if the model always output the same wrong class
- **Accuracy Percentage Reward:** discrete reward to foster perfect answers
 - 0 if the answer is wrong
 - 1 if the answer is entirely correct

1.4 Reaction Replacement

Task: Reaction Replacement (RxR)

Dataset Source	USPTO Reaction 1M; ¹
Input	Question: Which chemical reaction is correct? Choose from the following options: A. In the following reaction, the reagents are: <chem>Cc1ncc(C=O)n1C1CC1</chem>, <chem>CC(C)OC=C(Br)C=O</chem>, <chem>Cl</chem>, <chem>O=C(c1cc(N2CCNC2=O)ccc1F)N1CCCN(c2nccs2)CC1</chem> and the product is: <chem>O=Cc1cnc2n1CCCC2</chem>. B. In the following reaction, the reagents are: <chem>Cc1ncc(C=O)n1C1CC1</chem>, <chem>CC(C)OC=C(Br)C=O</chem>, <chem>Cl</chem>, <chem>N=C1CCCCN1</chem> and the product is: <chem>CNC(=O)CC1(O)CCCN(C(=O)c2cncc(F)c2)C1</chem>. C. In the following reaction, the reagents are: <chem>Cc1ncc(C=O)n1C1CC1</chem>, <chem>CC(C)OC=C(Br)C=O</chem>, <chem>Cl</chem>, <chem>N=C1CCCCN1</chem> and the product is: <chem>O=Cc1cnc2n1CCCC2</chem>. D. In the following reaction, the reagents are: <chem>Cc1cccc10CCC(=O)N1CCCC(c2ccn[nH]2)C1</chem>, <chem>CC(C)OC=C(Br)C=O</chem>, <chem>Cl</chem>, <chem>N=C1CCCCN1</chem> and the product is: <chem>O=Cc1cnc2n1CCCC2</chem>. Make sure to give your choice A, B, C, or D inside the <answer>...</answer> tags.
Target Molecular Reaction (Choice)	C
Output	<think> Let’s evaluate each option step by step to determine which one is correct. Option A: The reagent: <chem>Cc1ncc(C=O)n1C1CC1</chem> matches with the molecule <chem>Cc1ncc(C=O)n1C1CC1</chem>.
Answer	<answer> C </answer>

Table 4: Reaction Replacement experiment description.

The reaction replacement tasks challenges the model to understand chemical reaction concepts, validity and ability to detect subtle structural inconsistencies. By providing the model with four nearly identical choices, chemical reaction notation coherence understanding is required. Each dummy reaction has one reagent randomly swapped, where starting from a correct USPTO reaction, we generate three “corrupted” variants by replacing a single reactant or product with the most Tanimoto-similar molecule drawn from a random batch of 50 Enamine50k compounds. In the prompt we provide the lists options A–D, each specifying reagent SMILES, conditions SMILES, and product SMILES, and the model is then instructed to answer one of the four choices as the correct one. The model is also instructed to think through each option step by step inside <think>...</think> and the answer is emitted inside <answer>...</answer> tags.

Reward Functions:

- **Continuous Format Reward:**
 - This reward is described in Section 3.2.1 in the Algorithm 3.
- **Accuracy Reward:**
 - 0 if the answer is wrong
 - 1 if the answer is entirely correct

1.5 Reaction Inversion

Task: Reaction Inversion (RxI)

Dataset Source	USPTO Reaction 1M; ¹
Input	Question: Which chemical reaction is correct? Choose from the following options: A. In the following reaction, the reagents are: <chem>BrCc1ccccc1</chem>, <chem>[K+]</chem>, <chem>[OH-]</chem>, <chem>O=C(O)c1ccc(OCc2ccccc2)cc1</chem> and the product is: <chem>CCOC(=O)c1ccc(O)cc1</chem>. B. In the following reaction, the reagents are: <chem>C=O</chem>, <chem>O=Cc1ccccc1</chem>, <chem>[B-]C#N</chem>, <chem>[Na+]</chem>, <chem>CN[C@H]1[C@@H](C)C[C@H](c2ccncc2NC(=O)OC(C)(C)C[C@H]1NC(=O)OC(C)(C)C</chem>, the conditions are: <chem>CO</chem>, <chem>[OH-]</chem>, <chem>[OH-]</chem>, <chem>[Pd+2]</chem>, and the product is: <chem>C[C@H]1C[C@@H](c2ccncc2NC(=O)OC(C)(C)C[C@H]1NC(=O)OC(C)(C)C)[C@H]1N</chem>. C. In the following reaction, the reagents are: <chem>CCOC(=O)C#N</chem>, <chem>CCOC(=O)Cl</chem>, <chem>Cc1ccoc1C=Nc1ccccc1</chem>, the condition is: <chem>C1(C)C(C)=CC=CC=1</chem>, and the product is: <chem>CCOC(=O)c1cc2ccoc2cn1</chem>. D. In the following reaction, the reagents are: <chem>CC1(C)OB(c2cn[nH]c2)OC1(C)C</chem>, <chem>Nc1nc(-c2cc3c(s2)-c2ccc(-c4cn[nH]c4)cc2OCC3)c(-c2ccccc2Cl)s1</chem> and the product is: <chem>Nc1nc(-c2cc3c(s2)-c2ccc(Br)cc2OCC3)c(-c2ccccc2Cl)s1</chem>. Make sure to give your choice A, B, C, or D inside the <answer>...</answer> tags.
Target Molecular Reaction (Choice)	C
Output	<think> Starting with option A: The reaction uses benzyl bromide <chem>BrCc1ccccc1</chem> ...
Answer	<answer> C </answer>

Table 5: Reaction Inversion experiment description

The reaction inversion task challenges the model to understand chemical reaction concepts, validity and ability to detect subtle structural inconsistencies. By providing the model with four completely different choices, strong chemical reaction notation coherence understanding is required. Each dummy reaction has one reagent randomly swapped with the longest string SMILES among the products, enabling us to obtain 4 different reaction choices. In the prompt we provide the lists options A–D, each specifying reagent SMILES, conditions SMILES, and product SMILES, and the model is then instructed to answer one of the four choices as the correct one. The model is also instructed to think through each option step by step inside <think>...</think> and the answer is emitted inside <answer>...</answer> tags.

Reward Functions:

- **Continuous Format Reward:**
 - This reward is described in Section 3.2.1 in the Algorithm 3.
- **Accuracy Reward:**
 - 0 if the answer is wrong
 - 1 if the answer is entirely correct

1.6 Reaction True/False

Task: Reaction True/False (RxTF)

Dataset Source	USPTO Reaction 1M; ¹
Input	Question: Is this chemical reaction correct? In the following reaction, the reagent is: <chem>COC(=O)c1ccc(OC)c(OCCc2cccc(C#N)c2)c1</chem> , the conditions are: <chem>C1COCCO1</chem> , <chem>[Li+]</chem> , <chem>[OH-]</chem> , and the product is: <chem>COC1ccc(C(=O)O)cc1OCCc1cccc(C#N)c1</chem> .
Target Molecular Reaction Validity	True
Output	<think> First, I remember that <chem>LiOH</chem> , <chem>[Li+]</chem> . <chem>[OH-]</chem> is a strong base, so it's likely an acid-base reaction. The ester group in the starting material ...
Answer	<answer> True </answer>

Table 6: Reaction True/False experiment description

The Reaction True/False task is a binary derivative of the Reaction Replacement task. In this case, the model is asked to analyze and judge based on one single reaction, whether the reaction is correct or wrong. Each prompt presents one reaction—listing the reagent SMILES, the reaction conditions SMILES, and the product SMILES—and then asks “Is this chemical reaction correct?”. The examples are drawn from the Reaction Replacement set, where some of the reactions have been corrupted by swapping one random molecule in the reaction string by a new candidate. The model is instructed to reason step by step inside <think>...</think>, then has to emit <answer>True</answer> or <answer>False</answer> accordingly. This format was designed to simplify the reaction replacement task by providing only a binary label choice, allowing us to not only reduce the task complexity but also diminish the hallucination effects emanating from providing many examples in the prompt. In this scenario, the model only ever sees one reaction, and its ability to detect subtle mismatches in chemical transformations and to follow a true/false classification protocol with transparent chain-of-thought is tested.

Reward Functions:

- **Continuous Format Reward:**
 - This reward is described in Section 3.2.1 in the Algorithm 3.
- **Accuracy Reward:**
 - 0 if the answer is wrong
 - 1 if the answer is entirely correct

1.7 Chemical Formula Balancing Task (CeB)

Task: Chemical Formula Balancing Task (CeB)

Dataset Source	The Perovskite Dataset ⁶
Input	You are now acting as a skilled chemistry expert. Your goal is to identify and complete the missing compound(s) represented by [MASK] in the given chemical reaction. Reaction: $6 \text{ BaCO}_3 + 1 \text{ Na}_2\text{CO}_3 + 4 \text{ SiO}_2 == [\text{Mark}] \text{ Na}_2\text{Ba}[\text{Mark}](\text{Si}[\text{Mark}]\text{O}[\text{Mark}])(\text{SiO}[\text{Mark}])_2 + 7 \text{ CO}_2$. Analyze the equation carefully, considering the law of conservation of mass and stoichiometry. Ensure the total number of atoms of each element on both sides of the equation remains balanced.
Target Molecule (Material)	$6 \text{ BaCO}_3 + 1 \text{ Na}_2\text{CO}_3 + 4 \text{ SiO}_2 = 1 \text{ Na}_2\text{Ba}_6(\text{Si}_2\text{O}_9)(\text{SiO}_3)_2 + 7 \text{ CO}_2$
Output	----- <thinking>1. Element Count Analysis: - Reactants: - Ba: 6 (from 6 BaCO ₃) - Na: 2 (from Na ₂ CO ₃) - Si: 4 (from 4 SiO ₂) - C: 7 (6 from BaCO ₃ + 1 from Na ₂ CO ₃) - O: 6×3 (BaCO ₃) + 1×3 (Na ₂ CO ₃) + 4×2 (SiO ₂) = 29 - Products: - CO ₂ : 7 CO ₂ (7 C, 14 O) - Missing compound must account for remaining elements: 6 Ba, 2 Na, 4 Si, and 15 O (29 total O - 14 from CO ₂). 2. Structure of Missing Compound: - The formula template [MASK] Na ₂ Ba[Mark](Si[Mark]O[Mark])(SiO[Mark]) ₂ suggests: - 2 Na, 6 Ba (to balance Ba from reactants), 4 Si (from SiO ₂), and 15 O. - Silicate groups: (Si ₂ O ₉) and (SiO ₃) ₂ (to total 4 Si and 15 O). 3. Charge Balance: - Cations: Na ₂ (2+) + Ba ₆ (12+) = +14 charge. - Anion: Si ₄ O ₁₅ (4 Si × +4 + 15 O × -2 = +16 -30 = -14) balances the +14 charge. 4. Final Formula: - Na ₂ Ba ₆ (Si ₂ O ₉)(SiO ₃) ₂ consolidates all elements and charges correctly.</thinking>
Answer	<answer>6 BaCO ₃ + 1 Na ₂ CO ₃ + 4 SiO ₂ = 1 Na ₂ Ba ₆ (Si ₂ O ₉)(SiO ₃) ₂ + 7 CO ₂ </answer>

Table 7: Template for reporting RL experiments.

The goal of this task is to evaluate the performance of large language models (LLM) in predicting balanced chemical equations, specifically leveraging a reward mechanism to enhance the accuracy of chemical formula generation. The GRPO reinforcement learning data set was derived from a data set of chemical formulas focusing on perovskite compounds. The dataset was prepared by selectively masking certain products or their stoichiometric coefficients within balanced chemical equations.

Reward Function Design: The reward function combines two distinct components: format correctness and prediction accuracy.

- **Format Reward:**

- Initial reward is set to 0 if the generated chemical equation does not follow the correct format.
- An additional reward of 0.3 is given if the left-hand side (reactants) of the generated equation matches exactly with the ground truth.

- **Accuracy Reward:** This reward component assesses the similarity between the generated right-hand side (products) (y_{LLM}) and the ground truth (y_{GT}) using a predefined similarity metric, formulated as:

$$\text{reward} = \text{format_reward} + \text{sim}(y_{\text{LLM}}, y_{\text{GT}})$$

1.8 Conditional Material Generation (CMG)

Task: Conditional Material Generation (CMG)

Dataset Source	Materials Project ⁷
Input	Build a material that has O, Te, Tm. Please keep your reasoning as concise as possible. For example <material> A A B B B <sg12></material> where A, B refer to elements and <sg12> denotes the space group for example: <material> Pa In Tc Tc <sg225></material>.
Target Material Composition	Any valid composition made up of the elements O, Te and Tm with a valid space group number e.g. <material> O O Te Tm Tm Te <sg127></material>
Output	To combine O, Te, and Tm, a space group like sg127 (for a 1:1:1 ratio) could work. A possible structure could be: <material> O O Te Tm Tm Te <sg127></material> This fits the space group 127, which is a common space group for spacefilling structures.(Note: Space groups can vary, but this is a reasonable choice for a simple ternary compound.)
Answer	<material> O O Te Tm Tm Te <sg127></material>

Table 8: Conditional Material Generation experiment description

This task aims to leverage the scientific knowledge embedded in MiST-trained LLMs to generate novel materials from a specified set of elements. The experiment focuses on the model’s ability to understand three-dimensional atomic relationships within crystal structures and, based on that understanding, produce valid compositions. If the model can perform this task with high accuracy, it could significantly enhance the efficiency and cost-effectiveness of the material generation phase in the materials discovery process.

Reward Function Design: The quality of the generated composition is measured by the metrics: validity, precision and novelty. Validity is assessed using SMOG³ validity, which checks whether the generated composition adheres to fundamental chemical rules, such as charge neutrality. Precision measures the model’s ability to follow instructions and correctly include the specified elements. It is computed using the following equation:

$$\text{Precision} = \frac{|E_{pi} \cap E_{qi}|}{E_{pi}},$$

where E_{pi} is the set of elements specified in the i -th prompt and E_{qi} is the corresponding generated element¹⁰. The novelty of the generated composition was determined based on whether the composition was present within the materials project dataset or was previously generated by the model. Furthermore, to ensure the model provided its generated solution in a valid format, the reward function also checked that the generated composition was enclosed within the <material>...</material> tags and that the assigned space group number lies within the valid range of 1 to 230.

Therefore, the reward function used to train the LLM for the conditional material generation task was:

$$R = \alpha_1 \text{Validity} + \alpha_2 \text{Precision} + \alpha_3 \text{Novelty} + \alpha_4 \text{Format},$$

where the parameters α_i for $i = 1, \dots, 5$ are scaling factors for each portion of the reward.

1.9 Binary Compound Structure Relaxation Task (CrR)

Task: Binary Compound Structure Relaxation Task (CrR)

Dataset Source	Materials Project Trajectory (MPtrj) Dataset ⁴ , Materials Project ⁷
Input	Given a perturbed binary compound: formula Ti 1_int Ni 1_int, space_group_symbol 'P1' lattice_parameters a 2.74603000 b 2.94419691 c 2.94419691 alpha 119.99997545 beta 90.00000000 gamma 90.00000000 Zr 1_int 0.00000000 0.00000000 0.00000000 N 1_int 0.50000000 0.66670000 0.33330000 perform multiple steps Structural Relaxation on the given perturbed binary compound and reduce the internal energy. Please document your thought process within <think> </think> tags, and provide the final corrected structure in <answer> </answer> tags
Target Material	A more stable structure of the same material with lower inner energy.
Output	<thinking>The input structure describes ZrN in a hexagonal P-6m2 space group with significant initial distortions. The relaxation process improves stability through three key mechanisms: 1. Symmetry Restoration: The nitrogen position (0.5,0.6667,0.3333) deviated from standard WC-type coordinates (1/3,2/3,1/2). Relaxation shifts N to (0.3333,0.6666,0.5), aligning with proper Wyckoff positions for P-6m2. This restores three-fold rotational symmetry about [001], eliminating unnecessary symmetry-breaking distortions.2. **Bond Strain Reduction**: Initial Zr-N bonds (2.19 Å) were compressed relative to ionic radius predictions (Zr ³⁺ CN6: 0.85 Å+ N ³⁻ CN6: 1.32 Å= 2.17 Åideal). However, the metallic character of ZrN allows for shorter bonds. The relaxation increases bond length to 2.34 Å, balancing ionic size mismatch with metallic bonding contraction while reducing internal strain energy.3. **Lattice Parameter Optimization**: The original a-axis (2.746 Å) expanded to 3.15 Åpost-relaxation, better matching hexagonal close-packing requirements for Zr (atomic radius 1.60 Å). The c-axis remained near 2.94 Å, maintaining axial ratio c/a ≈ 0.93 - consistent with known WC-type structures rather than ideal HCP (c/a=1.633).The final structure shows improved prototype alignment with tungsten carbide (space group P-6m2), where both elements occupy high-symmetry positions. However, residual distortion in the pentagonal pyramidal coordination suggests possible metastability compared to the more common rock-salt ZrN polymorph. The increased symmetry and reduced bond strain would lower total energy, though the hexagonal phase may still lie above the cubic phase in energy.</thinking>
Answer	formula Zr 1_int N 1_int space_group_symbol 'P1' lattice_parameters a 3.15089295 b 3.15093266 c 2.93526192 alpha 90.00000000 beta 90.00000000 gamma 119.99958321 Zr 1_int 0.99999066 0.99998132 0.00000000 N 1_int 0.33330934 0.66661868 0.50000000</answer>

Table 9: Template for reporting RL experiments.

This task aims to utilize LLMs to perform preliminary relaxation (pre-relaxation) of unstable crystal structures, aiming to yield more stable structural configurations. This preliminary step is intended to substantially decrease computational costs and improve efficiency in subsequent high-accuracy Density Functional Theory (DFT) calculations. DFT calculations, while accurate, are computationally intensive. By leveraging LLM-generated pre-relaxation adjustments, the experiment seeks to effectively reduce the quantity of computationally unfavorable structures, thereby streamlining and accelerating the DFT computational pipeline.

Format Reward:

$$R_{\text{format}} = \begin{cases} -1, & \text{if } S_{\text{gen}} \text{ is valid Mat2Seq format and have lower inner energy than input structure} \\ -5, & \text{if } S_{\text{gen}} \text{ is valid Mat2Seq format} \\ -10, & \text{otherwise} \end{cases}$$

2 Benchmarking procedure

In this section we elaborate on the methods used to evaluate the models in the multiple ways displayed in Table 18. Here we give details of how diagnostic metrics have been computed (SCS, CCS), which evaluate the capabilities in LLMs that are necessary for success on chemical tasks in an RL setting. Additionally, performances on downstream tasks have been computed using benchmarks derived from each task (see Appendix above), along with different prompting techniques, that mark the difference between direct answer, or reasoning answer.

2.1 Latent Symbolic and Chemical Knowledge

2.1.1 Symbolic Competence Score benchmark

The Symbolic Competence Score benchmark measures the model’s latent capability to read and write correct chemical symbols. In this benchmark we focus particularly on SMILES, as organic chemistry spans a majority of our tasks. For this we collected 10000 valid SMILES from PubChem[?], such that no overlap exists with the MiST data. A second dataset is created with corrupted smiles based on these smiles, where corruptions are minimal, however render the smiles invalid. The corruption procedure is specified in Algorithm 1. The algorithm removes a random subset of key structural grammar elements (ring/branch brackets and digits) from the SMILES string, producing broken or ambiguous strings. Corruption rate ρ controls the proportion of removed elements, which for all our experiments has been set to 0.2.

Algorithm 1: SMILES Grammar Element Corruption

Input: SMILES string s , corruption rate ρ

Output: Corrupted SMILES string s_{corrupt}

```
1  Let  $\mathcal{G} = \{ (, ), [, ], 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 \}$  (grammar elements);
2   $L \leftarrow$  length of  $s$ ;
3   $I \leftarrow$  indices of  $s$  where  $s_i \in \mathcal{G}$ ;
4  if  $|I| = 0$  then
5  |   return  $s$ ;
6  end
7   $N_{\text{remove}} \leftarrow \max(1, \lfloor \rho \cdot |I| \rfloor)$ ;
8  Randomly select  $R \subseteq I$  with  $|R| = N_{\text{remove}}$ ;
9   $s_{\text{corrupt}} \leftarrow$  empty string;
10 for  $i \leftarrow 1$  to  $L$  do
11 |   if  $i \notin R$  then
12 | |   Append  $s_i$  to  $s_{\text{corrupt}}$ ;
13 |   end
14 end
15 return  $s_{\text{corrupt}}$ ;
```

Finally, evaluation happens in two stages. First, the log-likelihoods are computed using the model for the following string, that provides context for the string to look more natural:

The molecule represented with the SMILES [BEGIN_SMILES] smiles
[END_SMILES]

Where smiles is replaced by both the correct, and the incorrect SMILES string. The log-likelihoods corresponding to the smiles tokens are isolated by dropping the computed likelihoods associated with the context shown above. The two corresponding strings are thus

Original SMILES:

The molecule represented with the SMILES [BEGIN_SMILES]
O=C(O)C[C@H](O)C[C@H](O)CCn2c(c(c(c2c1ccc(F)cc1)c3ccccc3)C(=O)Nc4cccc4)C(C)C
[END_SMILES]

Corrupted SMILES:

The molecule represented with the SMILES [BEGIN_SMILES]
O=C(O)C[C@H](O)C[C@H](O)CCn2c(c(c(c2c1ccc(F)cc1)c3ccccc3)C(=O)Nc4cccc4)C(C)C
 [END_SMILES]

Average loglikelihoods are computed for the whole sample of 10000 SMILES in this manner, and SCS score is computed as the Cohen’s d effect size between the distributions of loglikelihoods of correct smiles, vs that of corrupted smiles.

Note that although the structure of material compositions is different from that of SMILES, the corruption method is similar, as key structural elements such as the space group number tag (<sg12>) and elemental symbols are replaced with special characters.

2.1.2 Chemical Competence Score benchmark

The Chemical Competence Score (CCS) evaluates a model’s latent ability to distinguish between chemically accurate and inaccurate factual statements. To construct this benchmark, we selected 1,000 samples from the test split of the SMolInstruct Molecule Description dataset¹¹, which was never used in all post-training stages. Each sample in the dataset consists of a brief description of an organic molecule. For example, one entry describes an acetamide as:

N-[4-(1,3-thiazol-2-ylsulfamoyl)phenyl]acetamide is a sulfonamide that is benzenesulfonamide substituted by an acetylamino group at position 4 and a 1,3-thiazol-2-yl group at the nitrogen atom. It is a metabolite of sulfathiazole. It has a role as a marine xenobiotic metabolite. It is a sulfonamide, a member of acetamides, and a member of 1,3-thiazoles.

For material data, we utilized Robocrystallographer⁵ to generate 600 natural text descriptions for crystal structures from the Material Project⁷. Here is an example entry:

AlN is Wurtzite structured and crystallizes in the hexagonal P6₃mc space group. Al(1) is bonded to four equivalent N(1) atoms to form corner-sharing AlN₄ tetrahedra. There are three shorter (1.90 Å) and one longer (1.91 Å) Al(1)-N(1) bond length. N(1) is bonded to four equivalent Al(1) atoms to form corner-sharing NAl₄ tetrahedra.

To create a contrastive benchmark, we generated an incorrect version for each entry by replacing one sentence in the original description with a sentence from a different one, while keeping the target molecule/crystal unchanged. Here is an example of an incorrect version of the above acetamide example with the edited section highlighted:

N-[4-(1,3-thiazol-2-ylsulfamoyl)phenyl]acetamide is a **tricyclic triterpenoid of the isomalabaricane group**. It is a metabolite of sulfathiazole. It has a role as a marine xenobiotic metabolite. It is a sulfonamide, a member of acetamides and a member of 1,3-thiazoles.

2.2 Task Benchmarks

The benchmarks have been obtained by selecting a subset of the datasets defined in Appendix 1, for each of the tasks.

2.3 Inference techniques

We observed that models’ full text generation often overflows the available context window, without providing any final answer within <answer> tags, thus preventing its correct evaluation. To overcome this, upon failure to generate an <answer> tag, we directly append the <answer> tag and retry the generation, biasing the model towards generating an answer at that point. Pseudo-code for this is provided in Algorithm 2.

Table 10: Evaluation methods for each reaction task

Task	Evaluation Method
Reaction Prediction (RxP)	Exact match with the groundtruth product
Reaction Naming (RxN)	Top-1 classification accuracy over the 10 reaction classes.
Reaction Replacement (RxR)	Multiple-choice accuracy (selecting the one correct reaction out of four).
Reaction Inversion (RxI)	Multiple-choice accuracy (selecting the one correct reaction out of four).
Reaction True/False (RxTF)	Binary classification accuracy (correct vs. incorrect reaction).

An extension of such an injection technique is that models can be biased from the beginning of the completion towards directly providing an answer, thereby allowing us to evaluate the effect of the intermediate text inside <think> tags. In Table 18 in the main manuscript, direct answer results are reported outside of the parentheses, while reasoning results are in parentheses.

Algorithm 2: Answer tag injection <answer> - Think and answer procedure

```

1 SetKwInOutInputInput
2 prompt, model_sampling_params, model, nbr_max_retries Output : A completion containing
   <answer>...</answer>
3 result ← llm.generate(prompt, sampling_params);
4 completion ← result.outputs[0].text;
5 for  $i \leftarrow 1$  to  $max\_retries$  do
   // Append the '<answer>' token to coax a proper tag
6   new_prompt ← prompt ++ completion ++ "<answer>";
7   result ← llm.generate(new_prompt, sampling_params);
8   complete_completion ← result.outputs[0].text;
9   if HasAnswer(complete_completion) then
10    return complete_completion;
11 return complete_completion ;                               // fallback if still no tag

```

3 Experimental settings

3.1 MiST: Mid-stage Scientific Training

Our MiST model is based on the Qwen-2.5-3B model. We continue the pre-training and perform SFT thereafter on a chemically enriched corpus spanning a diversity of sources, targeting the two prerequisites we proposed in the main manuscript.

The following configuration of hyperparameters was used for training:

Table 11: MiST Pretraining Hyperparameters

Parameter	Value
Model Architecture	Qwen-2.5-3B
Epochs	4 ($\sim 90,000$ steps)
Batch Size	32
Max/Min Learning Rate	$1 \times 10^{-5} / 1 \times 10^{-6}$
LR Warmup Steps	1,000
LR Decay Steps	1,000
Optimizer	AdamW
Loss Function	Cross-Entropy
Hardware	$32 \times \text{H100 GPUs}$
Total GPU Hours	640

After this stage, the model is further trained with SFT on instruction and Q&A data, as well as reasoning traces obtained from a stronger reasoning LLM, on more chemistry-relevant tasks; see the following section for more details. The following configuration was used:

Table 12: MiST SFT Hyperparameters

Parameter	Value
Model Architecture	Qwen-3B
Epochs	3 ($\sim 32,000$ steps)
Batch Size	32
Learning Rate	1×10^{-6}
Optimizer	AdamW
Loss Function	Cross-Entropy
Hardware	$32 \times \text{H100 GPUs}$

3.2 Reinforcement Learning experiments

The Open-R1 repository from Hugging Face (<https://github.com/huggingface/open-r1>) was forked and modified with additional features/optimizations for the GRPO experiments. Each training was run for 12 hours on four nodes (with four NVIDIA GH200 120GB GPUs), summing to 16 GPUs and 192 GPU-hours per training. The best hyperparameters are summarized in Table 13. A completion length of 8192 was used to let the model output long reasoning thoughts. The best hyperparameters and rewards were optimized using a total of 30k GPU-hours with variations in the experimental setups. The list of used rewards is described in Section 3.2.1.

parameter	value
per_device_train_batch_size	1
gradient_accumulation_steps	8
learning_rate	2e-6
lr_scheduler_type	cosine
warmup_ratio	0.03
beta	0.04
max_prompt_length	384
max_completion_length	8192
num_generations	8
use_vllm	true
vllm_max_model_len	8192

Table 13: Optimized hyperparameters used for the GRPO training experiments.

3.2.1 Rewards

The rewards designed for our GRPO experiments are grouped into two main categories:

- **Format reward:** the goal is to ensure that the trained model uses the appropriate format with reasoning (between <think> tags) and answer (between <answer> tags).
- **Accuracy reward:** the goal is to verify the answer of the model for the given task.

Accuracy reward: For the different tasks, different accuracy rewards are implemented in a continuous manner if possible. For SMILES-based tasks, the Tanimoto similarity score is generally used. However, for MCQA-based tasks, the rewards are usually discrete since the answers are correct or wrong. These rewards typically range from 0 to 1 (perfect answer).

Accuracy percentage reward: For each task, we also implement a discrete accuracy percentage reward to foster perfect answers and to log the training accuracy of the models. This reward is 0 if the answer is wrong and 1 if the answer is entirely correct.

Continuous format reward: A continuous format reward has been implemented with the structure described in Algorithm 3. The idea behind this reward is to output a score between -1 (very bad format)

and 1 (perfect format) with continuous steps to help the model with the learning of the expected format.

Algorithm 3: Incremental Formatting Reward Calculation

Input : Raw model output $o \in \text{String}$
Output : Formatting reward $r \in [-1, 1]$

```

1  $r \leftarrow 0.0$  // Initialize reward
2  $T \leftarrow \{<\text{think}>, </\text{think}>, <\text{answer}>, </\text{answer}>\}$ 
   // Check each tag appears exactly once
3 foreach tag  $\in T$  do
4   | if COUNT( $o$ , tag) = 1 then  $r \leftarrow r + 0.05$ 
5   | else  $r \leftarrow r - 0.05$ 
6 end

   // Check correct start and end tags
7 if STARTS_WITH( $o$ ,  $<\text{think}>$ ) then  $r \leftarrow r + 0.05$ 
8 else  $r \leftarrow r - 0.05$ 
9 if ENDS_WITH( $o$ ,  $</\text{answer}>$ ) then  $r \leftarrow r + 0.05$ 
10 else  $r \leftarrow r - 0.05$ 

   // Check think-answer boundary
11 if COUNT( $o$ ,  $</\text{think}>\backslash\text{n}<\text{answer}>$ ) = 1 then  $r \leftarrow r + 0.1$ 
12 else  $r \leftarrow r - 0.1$ 

   // Check answer block extraction
13  $m_1 \leftarrow \text{REGEX\_MATCH}(<\text{answer}>(.*)</\text{answer}>, o)$ 
14 if  $m_1 = \text{None}$  then
15   |  $r \leftarrow r - 0.2$ 
16 else if NUM_GROUPS( $m_1$ )  $\neq 1$  then
17   |  $r \leftarrow r - 0.05$ 
18 else
19   |  $r \leftarrow r + 0.2$ 
20 end

   // Check whole think \n answer pattern
21  $m_2 \leftarrow \text{REGEX\_MATCH}(<\text{think}>(.*)</\text{think}>\backslash\text{n}<\text{answer}>(.*)</\text{answer}>, o)$ 
22 if  $m_2 = \text{None}$  then
23   |  $r \leftarrow r - 0.4$ 
24 else if NUM_GROUPS( $m_2$ )  $\neq 2$  then
25   |  $r \leftarrow r - 0.1$ 
26 else
27   |  $r \leftarrow r + 0.4$ 
28 end
29 return  $r$ 

```

4 Data

4.1 Data sources and processing

4.1.1 FineWeb-Edu

The FineWeb-Edu can be found on Hugging Face (<https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu>)⁹. The subsets "CC-MAIN-2013-20" to "CC-MAIN-2024-10" were downloaded for a total of ~ 6 TB, which represents roughly 1.3T tokens and 1.26B individual texts. Based on the representative subset "sample-10BT" (also downloaded), the text sources were computed by taking the base URL (from the dataset column "url"), then these sources were sorted from the most prevalent to the least. We manually labeled the most prevalent sources as "chemistry", "non-chemistry", or "undetermined". The goal was to label a source as "chemistry" only if nearly all the texts from that source are about chemistry. On the other hand, a source is classified as "non-chemistry" only if there is no mention of chemistry in all the texts from that source. When a source contains a mix, like a school website with chemistry texts and texts for other fields, the label used is "undetermined", and the source is not used. After this manual labeling, the texts from "sample-10BT" were classified based on the labeled sources. It led to a ground truth of approximately 10,000 "chemistry" texts and 50,000 "non-chemistry" texts (out of the ~ 10 M texts found in "sample-10BT"). Based on this ground truth, a custom non-ML classifier was built using the word frequencies in "chemistry" and "non-chemistry" texts. The texts were lemmatized before building word frequency vectors for the two classes using a simple processing script that replaces any non-standard character with a space, before splitting the strings by the spaces. A custom vocabulary was also built to store these lemmatized texts in a tokenized manner. Other lemmatization methods (such as Spacy or NLTK) were also tried, but did not lead to better results and were extremely expensive to use on the full FineWeb dataset (>6 TB). After building the vocabulary and the word frequency vectors for the two classes, the formula below was applied to each FineWeb text to create an associated "chemistry score" (ranging from 0 to "infinity"). The frequencies of the lemma k in chemistry texts and non-chemistry texts are denoted f_k^c and f_k^n , respectively. The text chemistry score (TCS) is computed using the following equation:

$$TCS(text) := \frac{1}{N_{lemmas}} \sum_{\substack{k=\text{lemma} \\ \text{in text}}} w_k \quad \text{with} \quad w_k = \begin{cases} f_k^c/f_k^n, & \text{if } f_k^c/f_k^n > 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

This labeling strategy was applied to the entire FineWeb-Edu corpus, and the texts with $TCS > 4$ were retained, yielding a pretraining set of 1.4 billion tokens of high-quality chemistry-labeled texts. The threshold $TCS > 4$ was decided based on the PR curve plot shown in Figure 2. This threshold allows for high precision, and the quantity of texts retrieved was sufficient for our pretraining pipeline. Additional plots with the percentage of chemistry texts by threshold and the cumulative number of chemistry token counts by threshold can be observed in Figures 3 and 4, respectively. Some chemistry text examples (with their associated TCS scores) are shown in Figure 5.

4.1.2 PubChem

The first three million compounds from the PubChem database⁸ (CID from 1 to 3,000,000) were dumped using the PUG REST API with batched requests in October 2024. Each record contains these columns (among others): CanonicalSMILES, IsomericSMILES, IUPACName, and InChI. Since the molecule canonicalization algorithm used in the PubChem database is not the same as the one used by RDKit, all the compounds were re-canonicalized. The canonical SMILES consistency was also ensured for each compound by computing four canonical SMILES for each molecule:

- CanonicalSMILES $\rightarrow$ canonicalized using RDKit.
- IsomericSMILES $\rightarrow$ canonicalized using RDKit.
- IUPACName $\rightarrow$ SMILES using py2opsin and then canonicalized using RDKit.
- InChI $\rightarrow$ canonical SMILES using RDKit.

Then the four newly generated canonical SMILES were compared, and if a mismatch is found, the compound is discarded. This method filtered out approximately 40% of the compounds, and the duplicated canonical SMILES were also discarded. For the remaining compounds, four "SMILES

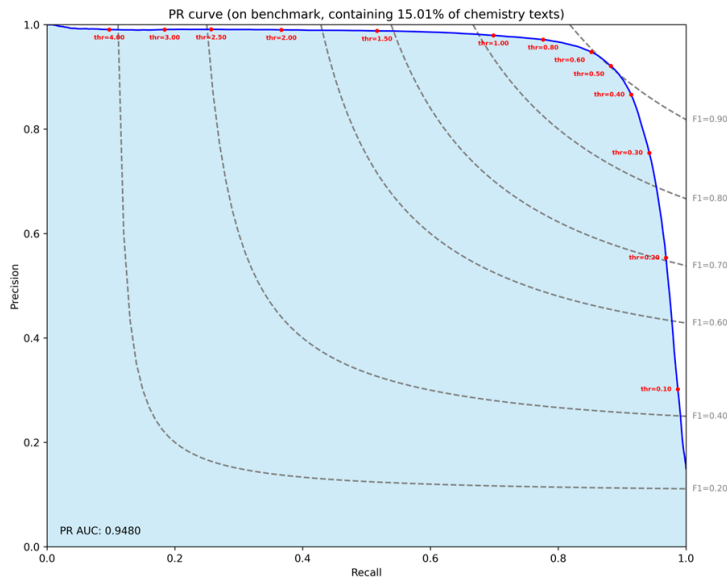


Figure 2: Precision-recall curve of the estimated retrieved chemistry texts based on the manually labeled ground truth. The different *TCS* thresholds are shown in red dots on the PR curve.

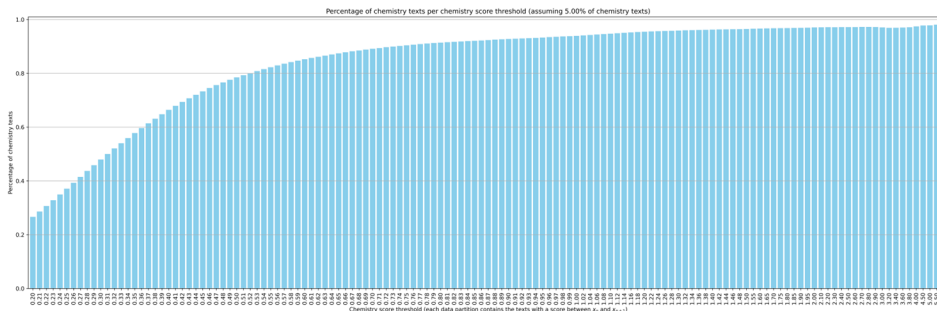


Figure 3: Estimated percentage of chemistry texts by *TCS* threshold.

variants" were computed using RDKit based on the canonical SMILES to have four non-canonical SMILES in each record. At the end of this processing script, an approximate of 1,800,000 compounds were kept and ready to be used. The dataset was then split in the following manner: the first million compounds (CID from 1 to 1,000,000) were used for pretraining, the second million compounds (CID from 1,000,001 to 2,000,000) were used for GRPO training, and the third million compounds (CID from 2,000,001 to 3,000,000) were used as the test split for benchmarking. Each split contains $\sim 600,000$ valid compounds. Multiple derived datasets were also generated for the different chemical tasks used with GRPO training (explained in Section 4.2 below).

4.2 Chemical Tasks Data sources

All MCQA-derived tasks for GRPO training are built on the USPTO Reaction 1M dataset, and the I2S dataset was built using the PubChem dataset from Section 4.1.2:

Reaction Prediction (RxP)

- The USPTO-480K dataset² consists of approximately 480K organic reactions, divided into training and test splits.
- We retained only reactions with a single product, resulting in roughly 400K training samples and 38K test samples.
- The first 10K reactions from the training set are used to generate reasoning traces.

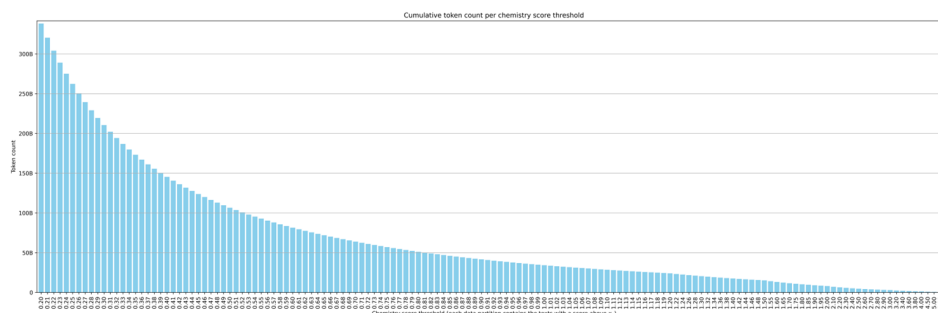


Figure 4: Estimated cumulative chemistry token count by TCS threshold.

```

score=5.18 | chemistry_id=733904
HTML by Rhodium
The methods available for the methylenation of aromatic o-dihydroxy compounds to the methylenedioxy compounds are unsatisfactory. The
methylenedioxy sulphate has been described, and was first prepared by Delapine by the action of paraformaldehyde on fuming sulphuric acid.
Methylenedioxy sulphate has now been prepared by a modification of the methods described by Delapine and Delaby in a yield of over 60%. It is
Reaction with o-Dihydric Phenols
Catechol reacts readily with methylene sulphate in 50% acetone solution in presence of sodium hydroxide to give catechol methylene ether.
To stirred molten fuming sulphuric acid containing 50% of sulphur trioxide (500 g), finely powdered paraformaldehyde (100 g) is added,
The methylenation of catechol in benzene containing potassium carbonate, or of its disodium salt in acetone. To a mixture of catechol (6
Protocatechuic aldehyde (1.4 g) in alcohol (10 ml) was treated with potassium hydroxide (4.9 g) in water (10 ml) and methylene sulphate
score=5.20 | chemistry_id=724531
An organic acid is a carbon-based compound with acidic properties, while an organic base is a carbon-based compound with basic pr
A typical acid is one that is a proton donor, while a base is usually a proton acceptor. Thus, by viewing it in this light, the a
pH = -log([H+])
However, since most organic acids are weak acids, then Ka and pKa are values that are also considered in acid-base calculations:
Ka = [H+][A-]/[HA]
While pKa = -log(Ka)
pH and pKa are also related in the following manner:
pH = pKa + log([A-]/[HA])
Thus, understanding the definition of an organic acid and base as well as the relationships between these two entities is crucial
score=2.10 | chemistry_id=812181
Analyzing the effects of varying hydrogen peroxide (H2O2) concentrations on the volume of its drop Jonah Jemina IB Chemistry SL 1
From the books, "Hydrogen Peroxide: Medical Miracle" by Douglass and "Applications of hydrogen peroxide and derivatives" by Jones
As a molecule, hydrogen peroxide has a double bent structure and is also a polar covalent particle. Due to these properties, hydro
To obtain different concentrations, 3% hydrogen peroxide will be diluted with varying amounts of distilled water. Each different
A drop of each solution will also be observed by analyzing their relative diameter and height sizes on a flat surface table using
Generally, the stronger the intermolecular forces are between molecules, the more they are attracted to one another resulting to
score=2.10 | chemistry_id=8114130
A DNA polymerase is an enzyme which makes DNA molecules from its nucleotide building blocks. DNA polymerases are essential for DNA repl
DNA polymerases also play key roles in other processes within cells, including DNA repair, genetic recombination, reverse transcription,
score=2.10 | chemistry_id=8114213
Brief History and Some Definitions
Liquid chromatography (LC) was defined in the early 1900s by the work of the Russian botanist, Mikhail S. Tswett. His pioneering studies fo
Tswett filled an open glass column with particles. Two specific materials that he found useful were powdered chalk [calcium carbonate] and
Tswett coined the name (chromatography) [from the Greek words chroma, meaning color, and graph, meaning writing-literally, color writing] to d
Figure A: Tswett's Experiment
Liquid chromatography (LC) Techniques
Liquid chromatography can be performed using planar [Techniques 1 and 2] or column techniques [Technique 3]. Column liquid chromatography is
Technique 1: The sample is spotted onto, and then flows through, a thin layer of adsorbent particles [stationary phase] fixed onto the
score=9.68 | chemistry_id=78642839
Should You Be Taking Zinc Tablets?
Zinc - a metallic element - occurs naturally in dairy foods, nuts, beans, red meat and many types of seafood. It plays an important role in
Symptoms of Zinc Deficiency
Maintaining a healthy level of zinc is important for your immune system and your ability to heal, as well as sustaining a healthy thyroid an
Zinc isn't a mineral people think too closely about. It simply does its job and isn't missed until a person is faced with a zinc deficiency.

```

Figure 5: Examples of labeled chemistry texts with the associated TCS scores.

- An additional 50K reactions, randomly selected from the remaining training data, are used for RLVF.
- A set of 500 reactions, randomly sampled from the test set, is used for benchmarking.

IUPAC to SMILES (I2S)

- The processed PubChem compounds (CID from 1,000,001 to 2,000,000) from the Section 4.1.2 are used as the base data.
- The canonical SMILES and the IUPAC were directly used from the dataset.

Reaction Naming (Rxn)

- Start from the full USPTO 1M reaction set.
- Use Rxn-Insight's class generation to detect the reaction name.
- Filter to 600000 samples, evenly distributing across the 10 classes.

Reaction Replacement (RrR)

- Duplicate each USPTO 1M reaction four times.
- For three copies, randomly select one molecule (reactant or reagent) to replace.

- Draw a batch of 50 candidate molecules from Enamine50k and compute Tanimoto similarity.
- Swap in the most similar molecule as the replacement.

Reaction Inversion (RxI)

- Take four instances of reactions in USPTO 1M, and invert one reagent with a product for 3 of them.
- The LLM is required to predict which one of the four reactions is still correct.

Reaction True/False (RxTF)

- Derived from the Reaction Replacement dataset.
- Present a single reaction (original or corrupted) and ask the model to judge its chemical correctness.

4.3 Material Tasks Data sources

Chemical Formula Balancing Task (CeB)

- A total of 1500 chemical formulas were selected from the Perovskite Dataset⁶ to form the data set, and the data set was then enhanced by randomly masking individual stoichiometric coefficients within products or entire product compounds using [MASK].

Conditional Material Generation (CMG)

- We selected 1000 samples from Materials Project⁷ and extracted the constituent elements from each sample to create our dataset. For example, the compound TeO₂ was decomposed into its constituent elements Te and O to form our training set.,

Binary Compound Structure Relaxation Task (CrR)

- We selected 2,000 binary compound crystal structures from the Materials Project⁷ across the following categories: Intermetallics, Semiconductors, Oxides, Sulfides, Nitrides, Carbides, Hydrides, Halides, Borides, Silicides, Phosphides, Arsenides, Tellurides, and Selenides. And we applied perturbations to alter the positions of certain atoms and modify the cell parameters of these structures to form our training dataset.

4.4 Resulting data mixture

The pretraining dataset was post-processed using an annotation pipeline to detect each molecule in the texts. For each molecule, the tags "[START_MOL]" and "[END_MOL]" were added to enclose it. Similarly, the SMILES were computed for each molecule and added between "[START_SMILES]" and "[END_SMILES]" tags after the molecule.

Table 14: MiST Pretraining Dataset Composition

Data Source	Tokens	Proportion
ChemRxiv + S2ORC	1.2B	41.37%
FineWeb (Q4–6)	1.4B	48.27%
PubChem Synthetic	120M	4.14%
Synthetic Reactions	100M	3.44%
CommonCrawl Replay	80M	2.75%
Total	2.9B	100%

Supervised fine-tuning was performed on the MiST - Qwen-3B model, primarily using chemistry-specific reasoning and instruction datasets, as follows:

Table 15: MiST SFT Dataset Composition

Data Source	Contents/Size
DeepSeek Rxn Traces	~7,000 samples
SmolInstruct	I2S, S2I, captioning, gen.
MMLU	350 general + 300 chemistry samples
Chain-of-Thought (CoT)	~27,000 samples

5 Compute resources

As described in Section 3.2 for the GRPO experiments, each training was run for 12 hours on four nodes (with 4 NVIDIA GH200 120GB GPUs or 8 AMD MI250x 128GB GPUs), summing to 16 GPUs and 192 GPU-hours per training. The best hyperparameters and rewards were optimized using a total of 30k GPU-hours with variations in the experimental setups. An additional 10k GPU-hours were used for the final runs, summing to a total of 40k GPU-hours.

6 Additional experimental results

6.1 MiST

We conducted other experiments to evaluate our MiST model’s performance on other tasks and in comparison with strong baselines from the literature. In particular, we compare against NatureLM¹⁰ and other general-purpose LLMs, on the task of SMILES to IUPAC and IUPAC to SMILES conversion. The results shown below put our MiST model (3B) on par with NatureLM 8B, while approaching the 8x7B MoE variant on IUPAC-to-SMILES conversion.

Table 16: Accuracy for IUPAC-to-SMILES and SMILES-to-IUPAC on benchmark datasets. The best value in each column is shown in bold.

Model	IUPAC-to-SMILES	SMILES-to-IUPAC
STOUT	0.735	0.565
GPT-4	0.033	0.0
Claude 3 Opus	0.177	0.0
LlaSMol_Mistral	0.701	0.29
NatureLM (1B)	0.476	0.284
NatureLM (8B)	0.679	0.517
<i>Qwen+MiST+SFT</i>	0.682	0.445

6.2 RL

From Table 17, it can be observed that the base model, Qwen-2.5 3B, possesses a degree of domain knowledge in materials science sufficient to generate some valid compositions. However, the relatively low scores suggest that the model is primarily retrieving compositions seen during training or generating valid combinations through rough heuristics. This is further supported by its low SCS, which indicates a limited understanding of compositions at the symbolic level.

The introduction of MiST leads to a significant improvement in SCS, as MiST specifically targets symbolic competence during training. However, since the model was not trained directly on materials science data and has a relatively small parameter size, it likely replaced some of its prior knowledge with representations more aligned with SMILES syntax. This shift contributes to the lower validity and precision scores, reflecting a reduced ability to follow instructions in non-SMILES-based tasks. As a result, the model often fails to generate outputs in the required format, especially when it encounters ambiguous prompts or reaches its maximum output length.

Fine-tuning the MiST model using SFT yields improvements in both SCS and instruction-following ability, as evidenced by higher validity and precision scores. These gains suggest that the model is able to recover some materials science knowledge while refining its symbolic understanding. However, the low novelty score indicates limited generalization, implying that the model is overfitting to training data and struggles to produce truly new compositions.

In comparison, SFT applied directly to the base Qwen-2.5 3B model results in high validity and precision but retains a poor SCS score. This contrast highlights that symbolic competence is primarily achieved through MiST, not SFT. Additionally, the low novelty score again suggests overfitting, as the model continues to rely on memorized examples rather than generating original compositions.

When combining MiST, SFT, and RL, there is a substantial improvement in novelty, indicating that the model is better able to utilize its symbolic understanding and domain knowledge to generate rather than recall compositions. This suggests that while base models have weak symbolic competence, MiST significantly enhances this capability. Though MiST initially reduces instruction-following ability due to longer and more complex outputs, SFT helps regain this ability for specific tasks. Ultimately, RL fine-tuning balances symbolic competence with domain-specific generation, enabling the model to produce valid, precise, and novel compositions using the specified elements.

In contrast to the findings observed in the Conditional Material Generation task, we did not detect any notable improvement in CCS after introducing MiST to the Binary Crystal Structure Relaxation task. This discrepancy arises because the Binary Crystal Structure Relaxation task specifically emphasizes

Table 17: CMG = Conditional Material Generation.

Model	SCS $\uparrow$	CCS $\uparrow$	Validity $\uparrow$	Precision $\uparrow$	Novelty $\uparrow$
Qwen-2.5 3B	0.122	0.828	58.6	68	74.8
+MiST	0.989	0.795	1.2	0.67	84.6
+SFT	1.142	0.785	34.8	38.5	49.2
+RL	0.893	0.777	73.8	97.1	91.3
Ablations					
no MiST + SFT	0.199	0.824	87.4	93.9	60.2

Table 18: CrR = Binary crystal structure relaxation, CeB = Chemical formula balancing.

Model	Metrics		Reasoning tasks	
	SCS $\uparrow$	CCS $\uparrow$	CrR $\uparrow$	CeB $\uparrow$
Qwen-2.5 3B	0.346	0.834	0	1.2
+MiST	0.355	0.795	0	26
+SFT	0.528	2.361	16.2	29.2
MatSci Tasks				
+RL(CrR)	0.447	2.599	65	—
+RL(CeB)	1.653	0.666	—	47
Ablations				
no MiST + SFT(CrR)	0.573	2.652	12.6	—
no MiST + SFT(CeB)	1.494	0.849	—	45

structural relaxation, a domain not directly targeted by MiST training. Consequently, MiST did not enhance the model’s chemical competence related to structural relaxation.

However, subsequent fine-tuning via SFT successfully incorporated relevant domain knowledge into the model, resulting in substantial performance improvements on the task. This step notably increased the model’s capability to accurately execute structural relaxations, which was previously limited. Moreover, further refinement through reinforcement learning (RL) effectively enhanced the model’s success rate, demonstrating that the integration of RL optimally balances domain-specific expertise with task-oriented performance improvements.

We further conducted an additional analysis across all 200 test set datapoints, and observed that the model performed comparably across the five crystal systems included in the test set.

Table 19: Summary of Crystal Systems for the MiST + SFT + RL (CrR) Model. This table presents a detailed breakdown of the performance (accuracy) of the MiST + SFT + RL (CrR) task, as shown in the Table, evaluated separately across different crystal systems.

Crystal System	Average Accuracy	Total Samples
Tetragonal system	0.6383	47
Orthorhombic system	0.6897	29
Hexagonal system	0.6250	72
Trigonal system	0.6572	35
Monoclinic system	0.7143	7
Cubic system	N/A	N/A
Triclinic system	N/A	N/A

We illustrate the capability of our Mist + SFT + RL model to reduce the inner energy of a perturbed, unstable ZnSe-P4_nmm crystal structure within 10 steps, where the stable state of the ZnSe-P4_nmm crystal has an inner energy of -2.94069766998291.

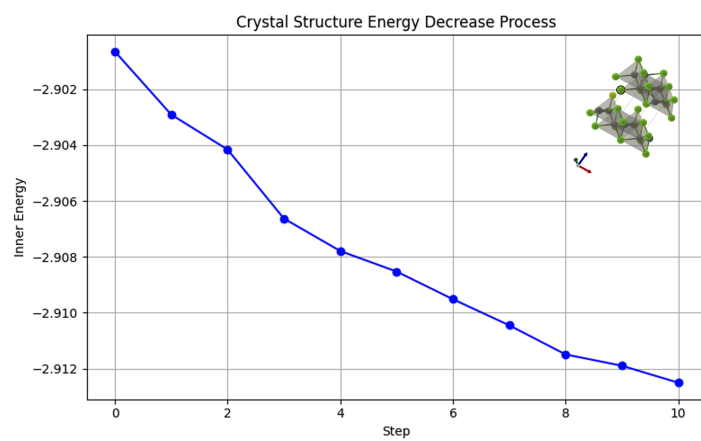
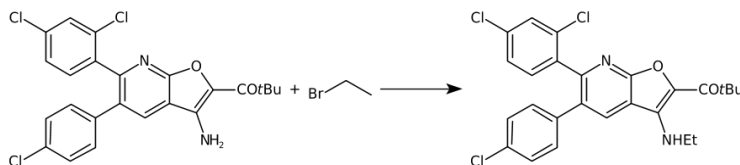


Figure 6: Graph demonstrating the relaxation of the ZnSe-P4_nmm crystal structure with the Mist + SFT + RL model

7 Model analysis

Here, we provide some analysis of the models' behavior. For the following example reaction:



Question

Reason and predict the correct product in

SMILES notation from the following reaction

CC(C)(C)C(=O)c1oc2nc(-c3ccc(Cl)cc3Cl)c(-c3ccc(Cl)cc3)cc2c1N.CCBr

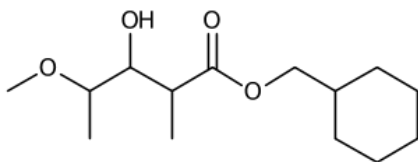
Correct answer

CCNc1c(C(=O)C(C)(C)C)oc2nc(-c3ccc(Cl)cc3Cl)c(-c3ccc(Cl)cc3)cc12

After RL training, although the output format is correct, the base model Qwen2.5-3B (+RL) fails to predict the correct product of this reaction, with simple reasoning and incorrect reaction type (esterification reaction). On the other hand, the variant with SFT exhibits a more detailed reasoning that delves into the functional groups of the precursors. However, it hallucinates the presence of sodium hydride and the hydroxyl group, which are not present in the input reagents, leading to an incorrect result. This could be attributed to the lack of SMILES symbolic understanding. Lastly, the model that underwent the full pipeline (MiST + SFT + RL) successfully reasons the correct product with reasonable description of the reactants' structure and mechanism, showing better symbolic understanding and Chemical knowledge.

The full responses of these models are given below:

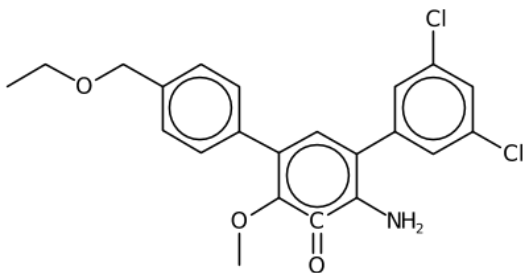
Qwen2.5-3B + RL [Wrong]



<think> The given reactants are a carboxylic acid and a bromoalkane. The reaction between these two compounds is an esterification reaction, where the carboxylic acid reacts with the alkyl halide to form an ester and a halide. The product of this reaction is an ester with the general formula R-COOR'. In this case, the product is a carboxylic acid ester with the general formula R-COOR', where R is the alkyl group and R' is the alkyl group from the bromoalkane.

</think>

<answer> COC(C)C(=O)C(C)C(=O)OCC1CCCCC1 </answer>



<think> The reaction involves the alkylation of a phenolic oxygen in the first compound using ethyl bromide CCBr under basic conditions. Here's the step-by-step mechanism:

1. Deprotonation**: Sodium hydride [H-].[Na+] ([Na+][H-]) deprotonates the hydroxyl group (-OH [OH-]) on the benzene c1ccccc1 ring of the first compound (CC(C)(C)C(=O)c1oc2nc(-c3ccc(Cl)cc3Cl)c(-c3ccc(Cl)cc3)cc2c1N), generating a phenoxide [O-]c1ccccc1 ion. The phenoxide [O-]c1ccccc1 is a strong nucleophile.

2. Nucleophilic Substitution(SN2)**: The phenoxide [O-]c1ccccc1 ion attacks the electrophilic carbon [C] in ethyl bromide CCBr. The bromide [Br-] ion (Br) is displaced, forming a new ether CCOCC bond (C-O-C) between the benzene c1ccccc1 ring and the ethyl group.

3. Product Formation**: The final product retains the original fused aromatic system of the first compound, with the hydroxyl group replaced by an ether CCOCC-linked ethyl group. The tert-butyl carbonyl (CC(C)(C)C(=O)), oxazole c1cocn1 ring (oc2...), and amine group (N) remain unchanged.

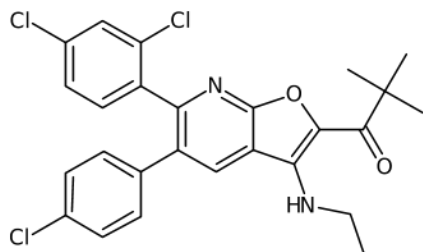
Product SMILES**: CCOCc1ccc(-c2cc(-c3cc(Cl)cc(Cl)c3)c(N)c(=O)c2OC)cc1

Explanation**: CCOC Ethyl ether CCOCC group formed from ethyl bromide CCBr. c1ccc(...)cc1: Benzene c1ccccc1 ring from the first compound. c2cc(...)c(N)...: Fused oxazole c1cocn1 ring(positions 1-6) with substituents retained. c3cc(Cl)cc(Cl)c3: Chlorinated benzene substituent from the first compound. N Amine group preserved in the fused ring. OC Oxazole c1cocn1 ring(positions 2-5) with a hydroxyl group replaced by the ether CCOCC linkage.

This reaction exemplifies a Williamson ether synthesis, where a phenoxide [O-]c1ccccc1 displaces a halide in the presence of a strong base to form an ether CCOCC. </think>

<answer> CCOCc1ccc(-c2cc(-c3cc(Cl)cc(Cl)c3)c(N)c(=O)c2OC)cc1
</answer>

Qwen2.5-3B + MiST + SFT + RL [Correct]



<think> The reaction involves the alkylation of a primary amine group in the first molecule by bromoethane CCB_r(CCB_r). Here's the step-by-step mechanism:

1. Reactant Analysis**:

First molecule**:

 A bicyclic compound with a tert-butyl group(CC(C)(C)C(=O)), a furan c1ccoc1-like ring(c1oc2nc(...)cc2c1N), and two chlorobenzene Clc1ccccc1 substituents(c3ccc(Cl)cc3 and c4ccc(Cl)cc4).

Second molecule**:

 Bromoethane CCB_r(CCB_r), where the bromine [Br] is a good leaving group.

2. Mechanism**:

The lone pair on the primary amine nitrogen [N](N [N]) in the first molecule acts as a nucleophile. This nitrogen [N] undergoes an S_N2 reaction** with bromoethane CCB_r. The amine attacks the electrophilic carbon [C] adjacent to the bromine [Br] in bromoethane CCB_r, displacing the bromide [Br⁻] ion(Br). The result is the formation of a secondary amine** in the first molecule, where the nitrogen [N] is now bonded to an ethyl group(CC).

3. Product Structure**:

The nitrogen [N] in the furan c1ccoc1-like ring is replaced by a CH₂CH₃ group, making it NCC** in the product. All other substituents(tert-butyl, chlorobenzenes [C-]#CCl, and furan c1ccoc1 ring) remain unchanged.

Final Answer**:

The product SMILES is

CC(C)(C)C(=O)c1oc2nc(-c3ccc(Cl)cc3Cl)c(-c3ccc(Cl)cc3)cc2c1NCC. The reaction selectively alkylates the primary amine group in the bicyclic compound with bromoethane CCB_r, converting it to a secondary amine while retaining the rest of the structure. </think>

<answer>

CC(C)(C)C(=O)c1oc2nc(-c3ccc(Cl)cc3Cl)c(-c3ccc(Cl)cc3)cc2c1NCC

</answer>

8 Reproducibility

All the code used to produce the results presented in this work can be found under <https://figshare.com/account/articles/29132657>. The continued pretraining and supervised fine-tuning, as described in Section 4 and Appendix 3, have been conducted using the nanotron library (see <https://github.com/huggingface/nanotron>). The configuration files and datasets used are released at <https://figshare.com/account/articles/29132657>.

Table of Released Assets

Asset	Usage Instructions	License/Citation Info	Location/URL
Source code	Download and unzip. See README.md for installation and experiment scripts (run_train.py).	MIT License. Please cite this paper.	https://figshare.com/account/articles/29132657
Model checkpoints	Download the archive. Full instructions in README.md.	MIT License. Please cite this paper.	https://figshare.com/account/articles/29132657
Datasets (pretraining/fine-tuning splits)	Download files; load as a HuggingFace Dataset.	For research use only. Cite the original dataset and this paper.	https://figshare.com/account/articles/29132657
Training configs	Config YAML files for nanotron available as .yaml; pass as argument to Nanotron CLI.	MIT License.	https://figshare.com/account/articles/29132657

Table 20: List of digital assets released with this work, including usage instructions and licensing/citation information. Note: All assets are hosted anonymously on Figshare for double-blind review.

All digital assets (code, models, data splits, and configs) are provided through anonymous Figshare links for double-blind review, as recommended by NeurIPS guidelines. After publication, these will be migrated to a permanent repository.

References

- [1] Connor W. Coley, Regina Barzilay, Tommi S. Jaakkola, William H. Green, and Klavs F. Jensen. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Central Science*, 3(5):434–443, May 2017. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00064. URL <https://doi.org/10.1021/acscentsci.7b00064>. Publisher: American Chemical Society.
- [2] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.*, 10:370–377, 2019.
- [3] Daniel W. Davies, Keith T. Butler, Adam J. Jackson, Andrew Morris, Jarvist M. Frost, Jonathan M. Skelton, and Aron Walsh. Computational screening of all stoichiometric inorganic materials. *Chem*, 1(4):617–627, 2016. ISSN 2451-9294. doi: <https://doi.org/10.1016/j.chempr.2016.09.010>. URL <https://www.sciencedirect.com/science/article/pii/S2451929416301553>.
- [4] Bowen Deng. Materials Project Trajectory (MPtrj) Dataset. *arXiv preprint arXiv:2302.14231*, 7 2023. doi: 10.6084/m9.figshare.23713842.v2. URL https://figshare.com/articles/dataset/Materials_Project_Trajectory_MPtrj_Dataset/23713842.
- [5] Alex M Ganose and Anubhav Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, 2019.
- [6] T. Jesper Jacobsson, Adam Hultqvist, Alberto García-Fernández, Aman Anand, Amran Al-Ashouri, Anders Hagfeldt, Andrea Crovetto, Antonio Abate, Antonio Gaetano Ricciardulli, Anuja Vijayan, Ashish Kulkarni, Assaf Y. Anderson, Barbara Primera Darwich, Bowen Yang, Brendan L. Coles, Carlo A. R. Perini, Carolin Rehmann, Daniel Ramirez, David Fairen-Jimenez, Diego Di Girolamo, Donglin Jia, Elena Avila, Emilio J. Juarez-Perez, Fanny Baumann, Florian Mathies, G. S. Anaya González, Gerrit Boschloo, Giuseppe Nasti, Gopinath Paramasivam, Guillermo Martínez-Denegri, Hampus Näsström, Hannes Michaels, Hans Köbler, Hua Wu, Iacopo Benesperi, M. Ibrahim Dar, Ilknur Bayrak Pehlivan, Isaac E. Gould, Jacob N. Vagott, Janardan Dagar, Jeff Kettle, Jie Yang, Jinzhao Li, Joel A. Smith, Jorge Pascual, Jose J. Jerónimo-Rendón, Juan Felipe Montoya, Juan-Pablo Correa-Baena, Junming Qiu, Junxin Wang, Kári Sveinbjörnsson, Katrin Hirslandt, Krishanu Dey, Kyle Frohna, Lena Mathies, Luigi A. Castriotta, Mahmoud. H. Aldamasy, Manuel Vasquez-Montoya, Marco A. Ruiz-Preciado, Marion A. Flatken, Mark V. Khenkin, Max Grischek, Mayank Kedia, Michael Saliba, Miguel Anaya, Misha Veldhoen, Neha Arora, Oleksandra Shargaieva, Oliver Maus, Onkar S. Game, Ori Yudilevich, Paul Fassel, Qisen Zhou, Rafael Betancur, Rahim Munir, Rahul Patidar, Samuel D. Stranks, Shahidul Alam, Shaoni Kar, Thomas Unold, Tobias Abzieher, Tomas Edvinsson, Tudur Wyn David, Ulrich W. Paetzold, Waqas Zia, Weifei Fu, Weiwei Zuo, Vincent R. F. Schröder, Wolfgang Tress, Xiaoliang Zhang, Yu-Hsien Chiang, Zafar Iqbal, Zhiqiang Xie, and Eva Unger. An open-access database and analysis tool for perovskite solar cells based on the fair data principles. *Nature Energy*, 7:107–115, 2022. doi: 10.1038/s41560-021-00941-3. URL <https://doi.org/10.1038/s41560-021-00941-3>.
- [7] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013. ISSN 2166-532X. doi: 10.1063/1.4812323. URL <https://doi.org/10.1063/1.4812323>.
- [8] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1059. URL <https://doi.org/10.1093/nar/gkae1059>.
- [9] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, October 2024. URL <http://arxiv.org/abs/2406.17557>. arXiv:2406.17557.

- [10] Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, Zekun Guo, Yeqi Bai, Pan Deng, Yaosen Min, Ziheng Lu, Hongxia Hao, Han Yang, Jielan Li, Chang Liu, Jia Zhang, Jianwei Zhu, Ran Bi, Kehan Wu, Wei Zhang, Kaiyuan Gao, Qizhi Pei, Qian Wang, Xixian Liu, Yanting Li, Houtian Zhu, Yeqing Lu, Mingqian Ma, Zun Wang, Tian Xie, Krzysztof Maziarsz, Marwin Segler, Zhao Yang, Zilong Chen, Yu Shi, Shuxin Zheng, Lijun Wu, Chen Hu, Peggy Dai, Tie-Yan Liu, Haiguang Liu, and Tao Qin. Nature language model: Deciphering the language of nature for scientific discovery, 2025. URL <https://arxiv.org/abs/2502.07527>.
- [11] Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*, 2024.