

---

# CHEMREASONER: Heuristic Search over a Large Language Model’s Knowledge Space using Quantum-Chemical Feedback

---

Henry W. Sprueill<sup>1</sup> Carl Edwards<sup>2</sup> Khushbu Agarwal<sup>1</sup> Mariefel V. Olarte<sup>1</sup> Udishnu Sanyal<sup>1</sup>  
Conrad Johnston<sup>3</sup> Hongbin Liu<sup>3</sup> Heng Ji<sup>2</sup> Sutanay Choudhury<sup>1</sup>

## Abstract

The discovery of new catalysts is essential for the design of new and more efficient chemical processes in order to transition to a sustainable future. We introduce an AI-guided computational screening framework unifying linguistic reasoning with quantum-chemistry based feedback from 3D atomistic representations. Our approach formulates catalyst discovery as an uncertain environment where an agent actively searches for highly effective catalysts via the iterative combination of large language model (LLM)-derived hypotheses and atomistic graph neural network (GNN)-derived feedback. Identified catalysts in intermediate search steps undergo structural evaluation based on spatial orientation, reaction pathways, and stability. Scoring functions based on adsorption energies and reaction energy barriers steer the exploration in the LLM’s knowledge space toward energetically favorable, high-efficiency catalysts. We introduce planning methods that automatically guide the exploration without human input, providing competitive performance against expert-enumerated chemical descriptor-based implementations. By integrating language-guided reasoning with computational chemistry feedback, our work pioneers AI-accelerated, trustworthy catalyst discovery.<sup>1</sup>

## 1. Introduction

The discovery of new catalysts requires one to identify the optimal combination of chemical descriptors (or properties)

---

<sup>1</sup>Pacific Northwest National Laboratory <sup>2</sup>University of Illinois Urbana-Champaign, <sup>3</sup>Azure Quantum, Microsoft. Correspondence to: Sutanay Choudhury <sutanay.choudhury@pnnl.gov>.

*Proceedings of the 41<sup>st</sup> International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

<sup>1</sup>The code and datasets discussed in this paper are available at <https://github.com/pnnl/chemreasoner>.

and use these descriptors to propose catalysts. However, such descriptors are only empirically understood, presenting a challenge for computational studies of catalysis. Typically, chemists actively reason to mentally search through reactants, catalysts, and operating conditions that enable more energy-efficient chemical conversions. However, as discussed by Nørskov et al. (2011), linking microscopic surface properties to macroscopic catalytic performance via chemical descriptors remains a barrier to descriptor-based catalyst search.

Large Language Models (LLMs) (Wei et al., 2022; Ouyang et al., 2022; Taylor et al., 2022; Lai et al., 2023; OpenAI, 2023) offer a new opportunity to realize such a data-driven autonomous search to accelerate scientific discovery. Our work aims to enhance natural language reasoning capabilities with quantum-chemical feedback to discover optimal catalysts for target reactions.

**The Challenge** However, reasoning about complex catalytic processes requires modelling across multiple modalities, extending beyond the capabilities of existing language models. This includes merging scientific concepts from literature and property prediction with 3D atomistic configurations. Determining the best catalyst is a multi-step process, requiring reasoning about multiple macroscopic properties. The first step involves identifying an optimal set of chemical descriptors (e.g. “resistance to poisoning”, “porosity”) which are relevant to the reaction in question. Formally, given a set of these important descriptors  $\mathbf{P}$ ,  $|\mathbf{P}| = n$ , we want to identify the optimal subset  $\mathbf{R} \subset \mathbf{P}$ ,  $|\mathbf{R}| = r$  to consider when suggesting new catalysts. This yields  $P_r^n = \frac{n!}{(n-r)!}$  possible permutations to reason over. The amount of reasoning required scales combinatorially with the number of available properties, necessitating an autonomous reasoner: an LLM. Using its knowledge of scientific concepts, the LLM both proposes important properties and proposes the best possible catalysts (from all possible catalysts) that have these properties. Pruning this large space of candidate catalysts requires reasoning about the complex microscopic interactions that occur between atomistic structures in 3D space based on macroscopic properties (Figure 2). Further, while simple reactions can be assessed via adsorption energies

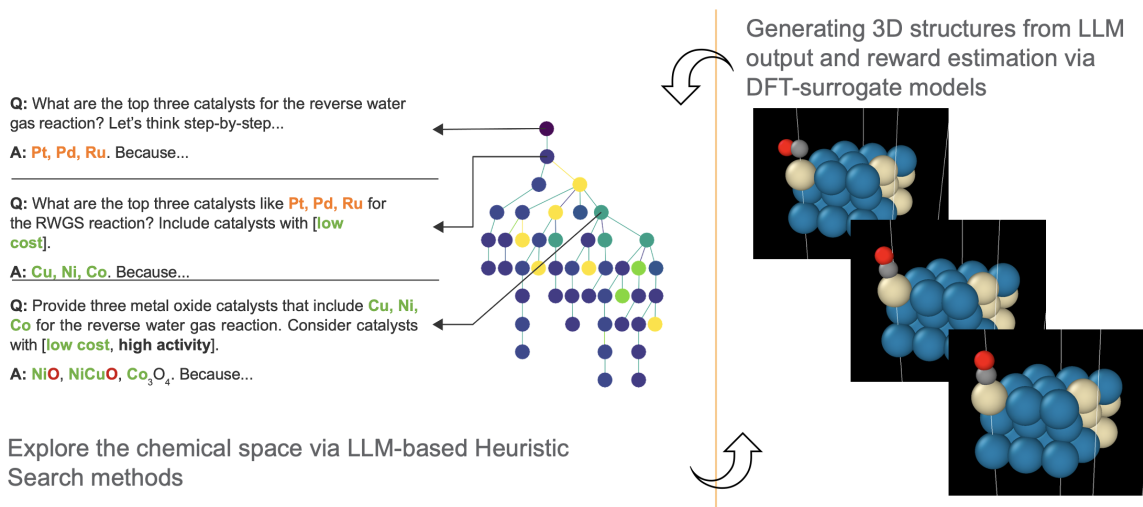


Figure 1. ChemReasoner successively “thinks” in terms of different constraints and factors, which are based on scientific principles and narrow down the set of possible candidates. It accomplishes that by prompting a language model with different combinations of chemical descriptors, yielding a tree-structured space of queries and potential candidates, and returns the optimal answer via efficient exploration of the search-space. ChemReasoner uses automated planning, based on previous reasoning, to initiate the exploration and guides it via a reward obtained from the exploration process to prune unpromising actions. We currently use “adsorption energy”, a key measure of reactivity as the reward function.

of 3D chemical structures, complex reactions demand consideration of multi-step reaction pathways and selectivity (Unsleber et al., 2023).

**Technical Approach** To solve this challenge, we propose a framework that combines LLM-driven heuristic search for catalyst discovery with structure-based scoring from atomistic graph neural networks (GNNs) trained from quantum chemistry simulations for guidance (Fig. 1). This framework formulates catalyst discovery as an uncertain environment where an agent (the LLM) pursues energetically favorable catalysts based on computational chemistry feedback. In each search step, the agent plans its actions (Huang et al., 2022; Hao et al., 2023) by 1) automatically identifying the optimal set of properties to consider, 2) generating new search prompts based on the identified properties, and 3) executing the prompts using sophisticated instruction following (Ouyang et al., 2022). Catalyst candidates identified in each step of the search are transformed into 3D atomistic representations of the catalyst-adsorbate structure (Zitnick et al., 2020). These representations enable evaluation via structural evaluation – including spatial orientation, energy barriers over reaction pathways, and stability – yielding a reward for catalyst suitability. This reward drives the LLM towards catalysts which enable reactions with minimal external energy, a crucial step for developing environmentally friendly industrial processes.

In this work, we make the following key contributions:

1. We introduce CHEMREASONER: a novel hypothesis

generation and testing framework unifying heuristic search over an LLM’s knowledge-space with quantum chemistry-guided feedback. This enables natural language-based reasoning for catalyst discovery with stronger domain guarantees obtained from computational chemistry methods.

2. We demonstrate the decisive impact of planning methods in automatically navigating chemical search spaces over a SOTA LLM-based implementation. Our purely LLM-planned approach with zero human input (CHEMREASONER-Planner) surpasses search guided by expert-selected chemical descriptors (CHEMREASONER-Expert) for two out of three categories in our evaluation benchmark.
3. Third, and uniquely, we establish the domain-grounding of language models via quantum chemical property feedback. We go beyond screening catalysts on adsorption energies alone and propose a methodology to reason in terms of reaction pathways and energy barriers.

Our work pioneers an AI-guided approach to computational catalyst screening and discovery. To facilitate community adoption and advancement of this extremely interdisciplinary and compute-intensive research, all of our datasets and code are freely available on github. This includes query benchmarks from catalysis experts, multi-modal provenance trails, over 700,000+ atomistic trajectories, and additional validation of catalyst candidates from density functional theory calculations.

## 2. Background and Related Work

### 2.1. Catalysis

Catalysts accelerate chemical reactions by lowering reaction barriers, without being consumed in the process. Heterogeneous catalysis, wherein the catalyst is in a different phase than the reactants and products, is widely used in industrial chemical processes (Dumesic et al., 2008). Developing novel heterogeneous catalysts with high activity and selectivity is essential to design energy efficient chemical process that paves a way towards sustainability (Zitnick et al., 2020; Hu & Yip, 2021; Mukhtar et al., 2022).

In heterogeneous catalysis, gases or liquids interact with a solid catalyst surface to enable a reaction (Greeley et al., 2002). This overall process consists of three elementary steps: 1) Adsorption - reactant molecules bind to the catalyst surface; 2) Surface reaction - adsorbed molecules react with each other to generate product(s); 3) Desorption - product(s) molecules desorb from the catalyst surface.

Adsorption energy of a reactant, reflecting its binding strength on a specific catalyst surface, is often identified as one of the key criterion for the activity of that catalyst. Fundamentally, adsorption energy of a reactant can be tuned by changing microscopic properties such as surface structure (crystal facet) or electronic configuration. Within catalysis, an overarching goal is to link these microscopic descriptors to macroscopic catalytic performance metrics to allow computation descriptor-based catalyst search (Nørskov et al., 2011). However, modeling complex catalyst surface reactions requires going beyond just adsorption descriptors (Xi et al., 2002). Interactions between reaction intermediates and competition between multiple possible reaction pathways must be considered as well (Kattel et al., 2017; Schwaller et al., 2019; Chen & Jung, 2022; Unsleber et al., 2023). With so many interactions in mind, we pursue a goal of developing models to reason compositionally about descriptors, structures, and pathways to generate high-quality hypotheses for potential catalysts, applied to the production of sustainable fuels.

### 2.2. LLMs for Chemistry

LLMs for chemistry can be divided into two categories: domain-specific models and adapted general-purpose models. Multimodal, domain-specific molecule-language models have recently emerged to target a number of problems in the chemistry domain (Edwards et al., 2021; Vall et al., 2021; Zeng et al., 2022; Xu & Wang, 2022; Su et al., 2022; Edwards et al., 2022a). For brevity, we discuss them further in Appendix D. On the other hand, general-domain LLMs such as GPT-4 (OpenAI, 2023) have been adopted in an agent-based approach to interface with chemistry-specific tools, allowing information gathering and hypothesis gener-

ation (Boiko et al., 2023; Bran et al., 2023). While this work is exciting, it differs from our approach, where we employ an overarching, domain-specific reward function to probe the LLM’s knowledge, enabling stronger, simulation-backed guarantees about complex scientific reasoning. We note that our method builds upon Sprueill et al. (2023). Unlike their approach, and the approach of Yao et al. (2023), which use LLM-computed rewards, we integrate true computational chemistry-based rewards to guide the model. Further, we integrate a context-aware planner into our search algorithm to automatically guide the search.

---

**Algorithm 1** Description of the CHEMREASONER framework.

---

**Require:** LLM, initial prompt  $P_0$ , number of children to generate  $N$ , number of children to keep  $M$ , target depth  $d$ . Initialize tree  $T$  with nodes  $P$  and edges  $(P, a_j)$ , scalar  $\gamma$ , stored values  $p(P, a_j)$ , and reward function  $R$ .

```

root(T) ← P0
Pcurr ← [root(T)]
Pnext ← []
for t = 1, . . . , d do
  for Pi ∈ Pcurr do
    Ai, p = P(Pi, p) ▷ Get action set and priors
    for aj ∈ arg top Nak ∈ Ai (p(aj)) do
      Pj* ← aj(Pi) ▷ Apply action aj
      T.append(Pj*)
      Pnext.append(Pj*)
    end
  Pcurr ← arg top MPj ∈ Pnext (R(Pj)) ▷ Calculate
  reward of LLM answer and downselect
end
return arg maxPj ∈ T (R(Pj))

```

---

## 3. System and Methods

This section elaborates algorithmic components of the architecture described in Algorithm 1. It can be broadly divided into two components: 1) LLM-planned and guided heuristic search over chemical space and 2) quantum-chemical feedback from graph neural network (GNN) models trained from density functional theory (DFT) simulations.

### 3.1. Heuristic Search

The goal of our heuristic search is to answer a user-specified natural language query by systematically exploring candidates from different regions in the chemical space. Heuristic searches have been applied for general problem solving and reasoning tasks for LLMs (Yao et al., 2023). Our work follows the catalysis-focused approach of Sprueill et al. (2023), where the original query (or prompt) and corresponding LLM answers are systematically modified by applying dif-

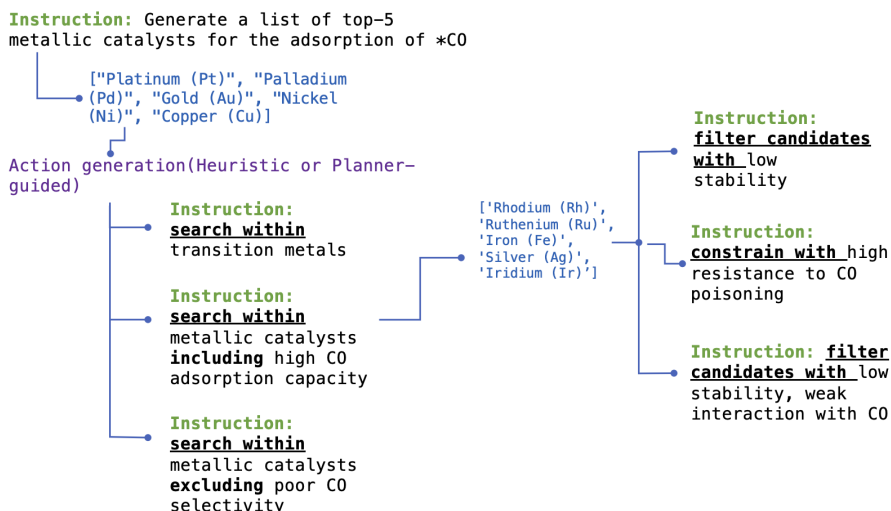


Figure 2. Illustration of CHEMREASONER search process (best viewed in color): The initial question generates base candidates, which are iteratively refined by adding an optimal set of constraints to the query and producing a new set of actions (or prompts) to explore the LLMs internal knowledge space. The optimal action set is chosen by 1) sampling from expert specified action space, or 2) automated generated by a planner component as illustrated in Figure 3. We describe the resultant structure shown as the “search tree” and each node in the tree represents a set of 3-tuple of (question, answer, reward). We refer to the initial query as a “root node.”

ferent screening criteria to iteratively contextualize the LLM prompts and answers into a narrower region of the chemical space. This process is illustrated in Figure 2, where the addition of chemical descriptors change the catalysts that are recommended. Since the number of branching pathways could hypothetically expand exponentially, we employ a domain-specific reward function to prune candidate catalysts that don’t show promise of high catalytic activity.

More formally, our goal is to search through chemical descriptors and design constraints to determine the optimal prompt, which leads the LLM to return the best candidate catalysts for a catalysis related query. Starting with a general prompt  $P_0$ , we use a set of actions to modify the prompt to improve the LLM output with respect to a reward function,  $R$ . Notably, CHEMREASONER-Planner generates its own action space  $\mathcal{A}$  (Figure 3).

**DEFINITIONS:** We define the *Search Tree* as a hierarchical tree consisting of (prompt, answer, and reward) nodes. Each node in this tree represents a state in the search space. Nodes are linked if an action  $a \in \mathcal{A}$  modifies the prompt from one node to the next. We denote a path from root to leaf node a *Reasoning Pathway*.

Following Sprueill et al. (2023), each node contains a template LLM prompt and an internal structured representation that provides additional context to the template prompt. This internal representation consists of 1) a natural language question, 2) an inclusion-exclusion list that includes or excludes specific chemical descriptors for target catalysts and 3) a relational operator that describes how the search can be shifted

from the previous query’s candidate catalysts to a different region in chemical space (i.e. similar-to or different-from the previous candidate catalysts). Starting the search with the root node, the search algorithm expands each node into a set of children nodes with a set of actions  $\mathcal{A}$ , which modify the internal representation of each node. The LLM then answers the modified prompt in each node, providing a set of candidate catalysts, and each candidate is scored using a reward function. Each layer of the search tree is pruned using a beam search algorithm a beam search algorithm (Rubin & Reddy, 1977), leaving only those nodes with the highest rewards. Finally, when the maximum search depth is reached, we select the node in the tree with the highest reward as the best answer to the initial prompt.

### 3.2. Planner-Guided Search

Our planner component is responsible for systematically expanding the search by contextually determining viable actions. Specifically, the planner selects actions based on the current context of the search, using vocabulary relevant to the current catalyst candidates. This contextual grounding automatically constrains the search direction in a scientifically coherent way. A technical description of the planner prompt is given in Appendix B.6.

Consider any node in the search tree where the planner generates the actions for a given node (shown in the orange boxes of Figure 3). Next, we execute the actions, creating several child nodes, and retrieve a set of top- $k$  candidate catalysts from the LLM (such as Cu, Pd etc.). As shown in

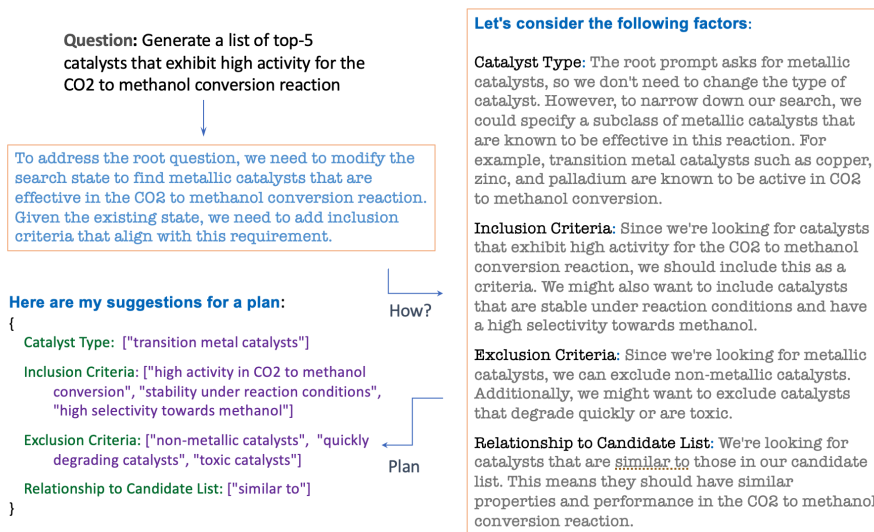


Figure 3. Planner-guided search action generation (best viewed in color): Given a query state defined by a question (shown in top-left) and the set of corresponding answers, the LLM is used as an optimizer to generate a “plan” for the next query. The LLM performs internal reasoning as shown in orange boxes. It accounts for the complete context from root query up to the current query node and generates a “query plan” with the attributes “catalyst type”, “inclusion criteria”, “exclusion criteria” and “relationship to current candidate list”.

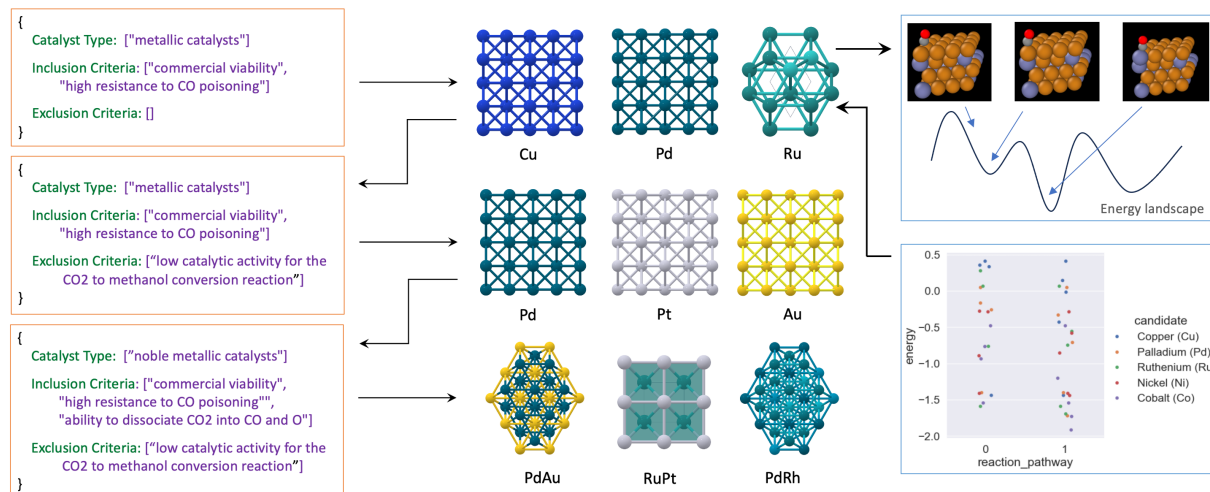


Figure 4. Illustration of planner guided heuristic search (best viewed in color) described in section 3.2 below. Note the systematic expansion of the query plan in the orange boxes (left column). The middle column shows illustration of 3D atomistic structures generated from chemical symbols. Each 3D structure is processed by a reward function that involves geometry relaxation and potentially deriving approximations of energy barriers in reaction pathways (right column). Visualizations of materials acquired from the Materials project structure finder (Jain et al., 2013).

figure 4, each of these candidates are then transformed into a 3D atomistic representation and evaluated by a reward function, which approximates catalytic activity (see sections 3.3 and 3.4 for details). All nodes at a given depth in the search tree are collected and only a subset of nodes are chosen for further search in the next iteration, filtered by highest reward. The process repeats iteratively (shown via rows in above diagram) until the maximum tree depth is reached.

Overall, by leveraging the LLM to contextually expand the tree search, CHEMREASONER-Planner explores candidate catalysts in a more interpretable, scientifically-grounded reasoning pathways. See section E for a complete trace of a planner guided search, using GPT-4.

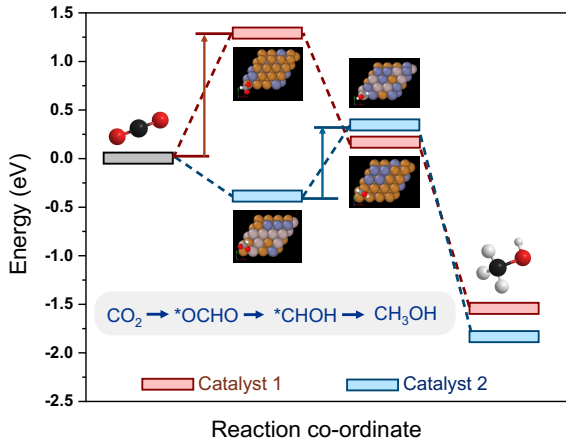


Figure 5. An illustration of reaction pathways corresponding to the conversion of  $\text{CO}_2$  to methanol for two different catalysts (red and blue bars shows the energy of different intermediates). The energy barrier (difference between the lower energy state and the higher energy state, described as hill climbing in the text) associated with these catalysts is shown by the red and blue arrows. For ease of comparison between the catalysts, the energies have been shifted such that both  $\text{CO}_2$  states have an adsorption energy of 0 eV.

### 3.3. Reward via Structure Optimization and Energy Prediction

Each reward function returns a real number as a measure of the catalyst’s goodness (higher is better) for a given input question. In this work, we implement two reward functions with different levels of complexity. The first reward function targets catalysts for the adsorption of particular chemical species, while the latter targets catalysts with higher approximate reactivity for certain reactions.

**Adsorption Energy-Based Reward:** To target catalysts for the adsorption of specific chemical species, we use an adsorption energy-based reward, which returns the adsorption energy of the most stable binding configuration of the catalyst as the reward. The computation begins with translating the symbolic representation of the catalysts (such as “Platinum”) and adsorbates (e.g., “\*CO”) into a 3D atomistic structure (Fig. 4 right). The stability and energy of a catalyst’s atomic structure directly impacts its catalytic activity and selectivity. Therefore, we compute the most stable configuration for the catalyst-adsorbate pair and use its adsorption energy as a measure of the reward (Section A). The optimization process, also known as the *relaxation process*, iteratively relaxes the atomic positions of the 3D structure until an energy minimum is found. A GNN (Gasteiger et al., 2021) is then used to calculate the adsorption energy from this state.

**Reaction Pathway-Based Reward:** This function measures the goodness of a catalyst considering multiple reaction pathways and intermediate steps. We initially obtained 5 possible reaction pathways from the LLM for the  $\text{CO}_2$  to

methanol/ethanol conversion reactions, as lists of chemical formulas. Since some of the pathways were redundant, we manually narrowed down the responses down to 2 reaction pathways for each application. These prompts were not re-evaluated for each tree search.

Given each reaction pathway, our reward function computes the adsorption energies for every intermediate step. Figure 5 shows two instances of the same reaction pathway for two different catalysts. As the figure shows, proceeding from one reaction step to another requires different amounts of energy, which depends on the catalyst and adsorbate interaction. Intuitively, moving from a lower-energy state to higher-one can be viewed as a “hill-climbing” in the energy landscape (indicated by the red and blue arrow for two different catalysts) and we formulate a function that would assign the highest reward to pathways with the smallest hills to climb (Eq. 1).  $\text{ads}_t$  is the intermediate at step  $t$  of the reaction and  $E_{\text{ads}_t}$  is the adsorption energy of  $\text{ads}_t$  on some catalyst. The overall reaction-based reward function for the top- $k$  catalysts is the average of  $r(c)$ , where  $c$  is one of the top- $k$  catalysts and  $r$  calculates the smallest maximum energy jump over all paths,

$$r(c) = - \min_{p \in \text{Paths}} \left( \max_{\text{ads}_t \in p} (E_{\text{ads}_t} - E_{\text{ads}_{t-1}}) \right). \quad (1)$$

## 4. Experiments

We conduct an experiment to evaluate if CHEMREASONER, using an LLM-guided heuristic search with quantum-chemistry feedback, can discover more novel and effective catalysts than state-of-the-art LLMs alone. Our experiments focus on three key research questions for enabling such a system.

- RQ1. Quantification of performance improvement:** Does heuristic-search guided by quantum-chemical feedback produce better catalyst candidates over querying state-of-the-art LLMs?
- RQ2. Characterization of key components:** What are the key parameters that control the computational complexity-system performance trade-off?
- RQ3. LLM Hypothesis testing:** How do we verify hypotheses generated by CHEMREASONER using domain knowledge? Which areas need further attention to make CHEMREASONER’s computational screening more accurate and interpretable?

### 4.1. Experimental setup

**Dataset** We conduct our experiments on an augmented version of a chemistry-focused reasoning query benchmark, originally proposed in (Sprueill et al., 2023), containing 145

	OpenCatalyst		BioFuels		CO <sub>2</sub> -Conversion	
	GPT-4	GPT-3.5	GPT-4	GPT-3.5	GPT-4	GPT-3.5
Chain-of-Thought	0.37	0.66	2.08	2.10	-0.62	-0.54
Self Consistency	0.73	0.76	2.08	2.12	-0.54	-0.36
CHEMREASONER-Expert	1.90	2.11	3.90	3.79	0.45	<b>0.78</b>
CHEMREASONER-Planner	<b>2.36</b>	2.16	<b>4.15</b>	3.29	0.01	0.49

Table 1. Final reward values of best recommended catalyst for each search variant. Our CHEMREASONER methods both significantly outperforms the GPT-4 baseline. Larger numbers are better and reflect GNN-predicted adsorption or reaction pathway-based rewards.

queries split between 3 general categories: OpenCatalyst, BioFuels, and CO<sub>2</sub>-Fuel. We adopt queries from Sprueill et al. (2023) for the first two categories and further enhance the dataset with the CO<sub>2</sub>-Fuel subset. First, the OpenCatalyst dataset is compiled from the set of adsorbates from the Open Catalyst Project 2020 dataset (Chanussot et al., 2010; Zitnick et al., 2020); it requires suggesting catalysts which each adsorbate strongly binds to (86 queries). Second, the BioFuels dataset is targeted at catalyst discovery for bio-fuel development (39 queries). These queries have been modified to target metallic catalysts, which is necessary for our reward calculation. Finally, we specifically target the conversion of CO<sub>2</sub> to methanol and ethanol (20 queries), a platform molecule, which can be used to produce fuels and chemicals for achieving net-zero carbon emissions (Ling et al., 2023; Mondal & Yadav, 2021). See section B for details on the queries in these datasets

**System Implementation** The LLMs used in our experimental setup included OpenAI GPT-3.5 and GPT-4. Although we initially benchmarked CHEMREASONER with Llama2 (Touvron et al., 2023), we found that its instruction-following capabilities in this domain were too limited to allow an evaluation.

For our GNN reward model, we utilized the GemNet-dT model (Gasteiger et al., 2021) from the Open Catalyst Project. Runtime configurations and inference scaling performance for each of these models are provided in section C.5. Inferences for both OpenAI models were executed in parallel using asynchronous execution features. The GNN inferences were run on a single DGX2/V100 or A100 GPU.

## 5. Towards Explainable Reasoning from Chemical Feedback

### 5.1. Reasoning Approaches

We evaluate two different variations of CHEMREASONER. CHEMREASONER-Expert is an implementation in which the action space is defined by catalysis experts. These actions (relation-operators and descriptors) are:

1. **Inclusion criteria:** high activity, high selectivity, low cost, novelty, low toxicity, high binding energy, high

conversion, high availability.

2. **Exclusion criteria:** low activity, low stability, low selectivity, low binding energy, high cost, high toxicity, low dispersion, low porosity, high scarcity, low conversion.
3. **Catalyst type:** metallic catalysts, monometallic catalysts, bimetallic catalysts, trimetallic catalysts.
4. **Relationship to previous candidate set:** include elements that are different from, include elements similar to, introduce new elements to, include elements from.

These actions are sampled with uniform probability, without using the same criteria twice. On the contrary, CHEMREASONER-Planner uses LLM-suggested actions for expanding the search space which do not require any expert specification. As such, CHEMREASONER-Planner’s actions can adapt to the changing context of the search, whereas CHEMREASONER-Expert’s action space remains static.

**Better LLM Translates into Search Efficiency** As Table 1 shows, both implementations of CHEMREASONER significantly outperforms the GPT-4 baseline. This is also visualized by Figure 8 (section C.3). Specifically, CHEMREASONER-Planner coupled with GPT-4 performs best for the OpenCatalyst and Biofuels query categories, whereas CHEMREASONER-Expert performs best for CO<sub>2</sub>-Conversion queries. As noted in Section 5.4 and shown in Table 2, the top-1 prediction of CHEMREASONER-Expert has high similarity with the current commercial catalyst for methanol synthesis (Etim et al., 2020). We also computed the average depth of the node containing the best answer for both variants of CHEMREASONER (Table 4). Lower average depth would indicate that our system is finding the best results more efficiently, and we observe that using GPT-4 leads to a reduction of 11.28% in the average search depth. The impact is more pronounced for CHEMREASONER-Expert than CHEMREASONER-Planner, which already obtains performance boost through the algorithmic contribution of planning.

**LLM’s Alignment with Reward function is key** The strong performance of the CHEMREASONER-Expert on CO<sub>2</sub>-Conversion queries is noteworthy, especially considering it is based on GPT-3.5-turbo. We hypothesize that the performance is related to a complex reward function.

For queries associated with the adsorption energy based reward function (OpenCatalyst and Biofuels) the LLM’s notion of a good catalyst typically aligns with lower adsorption energy (higher reward) profiles. Therefore, the planner effectively uses the LLM as an optimizing function for searching towards energetically favorable catalysts. However, the LLM’s notion of a good catalyst may not directly align with the complex reaction pathway-based reward function associated with CO<sub>2</sub> conversion. In general, it suggests that fine-tuning the LLM using a methodology similar to RLHF (Ouyang et al., 2022) may be a promising path for downstream tasks with complex reward functions.

## 5.2. Performance Characterization of Key Components

CHEMREASONER performs a large number of LLM and GNN inferences that influence both its performance and throughput. The following factors control the computational complexity of the overall execution.

**LLM Inference and Tree Search** The runtime of CHEMREASONER is  $O(N_{tree})$ , where  $N_{tree}$  is the number of nodes in the search tree.  $N_{tree}$  increases exponentially with the maximum depth (set to 5) and the branching factor for expanding each node in the search tree, denoted as  $N_{actions}$  (set to 8). We use a beam width parameter (set to 6) to control the number of nodes that are expanded in each iteration of the search. CHEMREASONER performs LLM inferences for each node in the search tree to 1) plan actions, 2) suggest candidate catalysts, and 3) transform LLM answers into atomistic structures. It is important to note the strong intra-node and inter-node dependencies between these LLM queries, as the output from one query is fed into the next ones. This dependency exists both within a node and to its children in the search tree. Therefore, robustness of LLM instruction-following behavior, specifically the ability to return answers in a consistent format, is a critical factor for an effective system. Overall, CHEMREASONER-Expert and CHEMREASONER-Planner executes  $2N_{tree}$  and  $3N_{tree}$  LLM inferences respectively.

To execute the entire benchmark of 145 queries with  $N_{max} = 300$  requires execution of 28,000-42,000 LLM inferences, which makes the the scalability and throughput of a generative LLM a critical factor for successful experimentation. By default, we batch LLM inferences in sizes of 48 and execute these inferences asynchronously for scalability. These two factors, namely the robustness of instruction-following and high-throughput execution of batched queries restricted our experiments to only GPT-3.5 and GPT-4 models. Our experiments with LLama2 were both affected by frequent spurious answers and relatively low throughput on A100 systems. We recognize the importance of the development and benchmarking of open-source models, but this particular topic should be the subject of

future work.

**GNN Inference and Reward Estimation** GNN inferences play an equally critical role for CHEMREASONER. A single execution of the adsorption energy-based reward function requires multiple inferences. Given a string-based representation of a catalyst, we first randomly initialize a 3D atomistic representation and oversample the structural conformation to generate  $N_{structs}$  samples (see Appendix A). Next, relax each of these  $N_{structs}$  structures (set to 16 by default) for a maximum specified iteration limit of  $N_{relax}$  (set to 64 by default), or until  $f_{max}$ . Therefore, each execution of each adsorption energy-based reward function requires  $N_{structs} * N_{relax}$  GNN inferences. For scalability, we group the GNN inferences using a batch size of 40. In reaction pathway-based reward computations, we perform this computation for every intermediate state in each pathway. Therefore, with  $N_{pathway}$  denoting the number of pathways and  $N_{rstep}$  denoting the average number of intermediate steps per pathway, a single reward estimation involves  $N_{structs} * N_{relax} * N_{pathway} * N_{rstep}$  GNN inferences. Given our CO<sub>2</sub>-fuel conversion queries which involve 2 reaction pathways comprising 4-5 steps, we execute 9,216 GNN inferences for every reaction-pathway based reward. On a single A100 or V100 GPU, batches can be evaluated in approximately 0.5 seconds.

## 5.3. Considering DFT-Based Rewards

A critical component for a reasoning system is accurate feedback. For this reason, the use of DFT-based rewards would be the gold standard reward calculations. However, due to their significant computational cost, we are forced to employ GNNs for high-throughput evaluation of catalyst candidates (Zitnick et al., 2020). We evaluated our top GNN results using DFT simulations; overall, we found that DFT calculation-based rewards diverge in some cases significantly from GNN predictions (Table 3). Here, we briefly examine the divergence between SOTA GNN models and ground truth DFT simulations.

One limitation of the GNN likely arises from the conversion of text-based catalyst recommendations from the LLM to 3D atomistic structures for the GNN calculation. Our structure generation method (see Section A) is limited to pre-defined lattice structures, (e.g., face centered cubic (FCC), body centered cubic (FCC), and hexagonal close packed (HCP)). While these lattice structures effectively describe mono-metallic structures, multimetallic compounds may exhibit more complex lattice structures. Thus, our assumptions of reference structures may not adequately reflect realistic catalyst structures. Furthermore, our multimetallic bulk structures were generated by randomly placing elements throughout the bulk, creating additional complexity due to the emergence of defects in the bulk structures.



While the GNN can predict energy for any given structures, the challenges mentioned above may lead to erroneous DFT calculations, which makes validation of our predicted catalysts difficult. Therefore, future research should explore novel methods to convert textual representations of catalysts into realistic 3D atomistic structures.

#### 5.4. CO<sub>2</sub> Hydrogenation: A Case Study

To evaluate the real-world efficacy of CHEMREASONER, we performed a literature evaluation of the top-5 predicted metallic catalysts for the conversion of CO<sub>2</sub> to methanol. These predictions from GPT-4, CHEMREASONER-Expert, and CHEMREASONER-Planner models are shown in Table 2. We find that the CHEMREASONER-Expert predictions include elements that make up the current commercial catalysts for methanol production Cu/ZnO/Al<sub>2</sub>O<sub>3</sub> (Etim et al., 2020). Cu, ZnO, and Cu-Zn sites were considered to generate the active sites for high methanol selectivity (Etim et al., 2020; Tisseraud et al., 2015).

On the other hand, CHEMREASONER-Planner predicted bimetallic precious metal alloys, which are well-known to catalyze hydrogenation reactions (Tawalbeh et al., 2023). The ensemble effect (Wang et al., 2019) in Pd-Au electrocatalysts was considered as key factor in achieving high activity for CO<sub>2</sub> reduction to methanol since Pd, Pt, Rh and Ru alone can shift the selectivity to methane (CH<sub>4</sub>) by uncontrolled hydrogenation of CO<sub>2</sub> (Pakhare & Spivey, 2014). Alloying allows for improved methanol selectivity compared to the performance of single metal catalysts (Tawalbeh et al., 2023; Etim et al., 2020; Tisseraud et al., 2015) that were predicted by GPT-4 and some by CHEMREASONER-Expert. Since both CHEMREASONER-Expert and CHEMREASONER-Planner predicted alloys, we consider their recommendations to have higher likelihoods of being successful catalysts.

## 6. Conclusion

We introduce a multi-modal framework unifying linguistic reasoning enabled by generative LLMs with atomistic structure based rewards, built on principles in catalysis and quantum chemistry. We demonstrate that such integration can enable the recommendations of catalysts that are critical for the development of energy efficient chemical conversion processes that combat climate change. Namely, CHEMREASONER was able to recommend CuAlZn for the conversion of CO<sub>2</sub> to methanol, the current commercially viable catalyst.

Our methodology makes use of instruction-following foundational LLMs, GPT-3.5 and GPT-4, avoiding the need for expensive finetuning. Additionally, our application of automatic planning methods removes the need to hard-code

domain specific knowledge into the search method. So long as one can define a systematic search space and a relevant domain specific reward function, one could adapt our method to other scientific applications in biology, chemistry, or materials science. While we recognize the importance of accessible, open-source LLMs, our experiments with Llama-2 revealed that such models may not currently possess the instruction following capabilities or scalability required for our heuristic search. Therefore, the development of open-source LLMs for scientific reasoning remains the subject of future work.

Finally, CHEMREASONER presents a new way to carry out computational investigations of catalyst materials with DFT surrogate models. Our application of a GNN based reward function grounds the LLM in reality, guiding the search towards catalysts with higher activity. Still, a limitation of CHEMREASONER is the reliance on existing reference structures, which may not accurately reflect real catalyst structures, providing a challenge to verification of results with DFT. This motivates the future development of more robust methods to convert textual representations of materials into 3D atomistic structures.

## Acknowledgements

This work is supported in part by the PNNL seed Laboratory Directed Research and Development (LDRD) program, Generative AI for Science, Energy and Security Lab-Wide Laboratory Directed Research and Development Investment.

This work is also in part based upon work supported by the Molecule Maker Lab Institute: an AI research institute program supported by NSF under award No. 2019897. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

PNNL acknowledges the Microsoft Azure Quantum Elements team for their assistance with this study and their contributions to this publication.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning for the chemical and material sciences. It involves the use of a large language model (LLM), which should be treated as a black-box. LLMs are known to “hallucinate”—produce erroneous or fabricated text—and may plagiarise other scientific work. The goal of this paper is to incorporate a domain specific reward function to mitigate hallucinations by the LLM by validating their answers

with computational chemistry feedback. Still, scientific results produced by LLMs should still be stringently verified by domain-experts to avoid false claims. The primary focus of this work is to discuss the strengths and weaknesses of the AI system proposed herein as a tool to accelerate scientific discovery, not necessarily to claim the discovery of a novel catalyst.

## References

- Artz, J., Müller, T. E., Thenert, K., Kleinekorte, J., Meys, R., Sternberg, A., Bardow, A., and Leitner, W. Sustainable conversion of carbon dioxide: an integrated review of catalysis and life cycle assessment. *Chemical reviews*, 118(2):434–504, 2018.
- Boiko, D. A., MacKnight, R., and Gomes, G. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.
- Bran, A. M., Cox, S., White, A. D., and Schwaller, P. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.
- Canakci, M. and Van Gerpen, J. Biodiesel production via acid catalysis. *Transactions of the ASAE*, 42(5):1203–1210, 1999.
- Cao, H., Liu, Z., Lu, X., Yao, Y., and Li, Y. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*, 2023.
- Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., et al. The open catalyst 2020 (oc20) dataset and community challenges. arxiv. *arXiv*, 2010.
- Chanussot\*, L., Das\*, A., Goyal\*, S., Lavril\*, T., Shuaibi\*, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, D., Zitnick, C. L., and Ulissi, Z. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 2021. doi: 10.1021/acscatal.0c04525.
- Chen, S. and Jung, Y. A generalized-template-based graph neural network for accurate organic reactivity prediction. *Nature Machine Intelligence*, 4(9):772–780, 2022.
- Cheng, A. H., Cai, A., Miret, S., Malkomes, G., Phielipp, M., and Aspuru-Guzik, A. Group selfies: a robust fragment-based molecular string representation. *Digital Discovery*, 2023.
- Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.
- Christofidellis, D., Giannone, G., Born, J., Winther, O., Laino, T., and Manica, M. Unifying molecular and textual representations via multi-task language modelling. *arXiv preprint arXiv:2301.12586*, 2023.
- Dal Corso, A. Pseudopotentials periodic table: From h to pu. *Computational Materials Science*, 95:337–350, 2014.
- Daza, Y. A. and Kuhn, J. N. Co 2 conversion by reverse water gas shift catalysis: comparison of catalysts, mechanisms and their consequences for co 2 conversion to liquid fuels. *RSC advances*, 6(55):49675–49691, 2016.
- Dumesic, J. A., Huber, G. W., and Boudart, M. Principles of heterogeneous catalysis. *Handbook of Heterogeneous Catalysis: Online*, 2008.
- Edwards, C., Zhai, C., and Ji, H. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.emnlp-main.47>.
- Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.26. URL <https://aclanthology.org/2022.emnlp-main.26>.
- Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-main.26>.
- Etim, U. J., Song, Y., and Zhong, Z. Improving the cu/zno-based catalysts for carbon dioxide hydrogenation to methanol, and the use of methanol as a renewable energy storage media. *Frontiers in Energy Research*, 8: 545431, 2020.
- Fabian, B., Edlich, T., Gaspar, H., Segler, M., Meyers, J., Fiscato, M., and Ahmed, M. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.
- Fang, Y., Liang, X., Zhang, N., Liu, K., Huang, R., Chen, Z., Fan, X., and Chen, H. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023.

- Gasteiger, J., Becker, F., and Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.
- Giannozzi, P., Baroni, S., Bonini, N., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Chiarotti, G. L., Cococcioni, M., Dabo, I., et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter*, 21(39): 395502, 2009.
- Giannozzi, P., Andreussi, O., Brumme, T., Bunau, O., Nardelli, M. B., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Cococcioni, M., et al. Advanced capabilities for materials modelling with quantum espresso. *Journal of physics: Condensed matter*, 29(46):465901, 2017.
- Greeley, J., Nørskov, J. K., and Mavrikakis, M. Electronic structure and catalysis on metal surfaces. *Annual review of physical chemistry*, 53(1):319–348, 2002.
- Grimme, S., Antony, J., Ehrlich, S., and Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of chemical physics*, 132(15), 2010.
- Guo, T., Guo, K., Nan, B., Liang, Z., Guo, Z., Chawla, N. V., Wiest, O., and Zhang, X. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *arXiv preprint arXiv:2305.18365*, 2023.
- Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.
- Hu, X. and Yip, A. C. Heterogeneous catalysis: enabling a sustainable future, 2021.
- Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.
- Huber, S. P., Zoupanos, S., Uhrin, M., Talirz, L., Kahle, L., Häuselmann, R., Gresch, D., Müller, T., Yakutovich, A. V., Andersen, C. W., et al. Aiida 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Scientific data*, 7(1):300, 2020.
- Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K. A. 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>.
- Kattel, S., Liu, P., and Chen, J. G. Tuning selectivity of co2 hydrogenation reactions at the metal/oxide interface. *Journal of the American Chemical Society*, 139(29):9739–9754, 2017.
- Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.
- Lai, T. M., Zhai, C., and Ji, H. Knowledge-enhanced biomedical language models. In *Journal of Biomedical Informatics*, 2023.
- Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Duřak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., Hermes, E. D., Jennings, P. C., Jensen, P. B., Kermode, J., Kitchin, J. R., Kolsbjerg, E. L., Kubal, J., Kaasbjerg, K., Lysgaard, S., Maronsson, J. B., Maxson, T., Olsen, T., Pastewka, L., Peterson, A., Rostgaard, C., Schiøtz, J., Schütt, O., Strange, M., Thygesen, K. S., Vegge, T., Vilhelmsen, L., Walter, M., Zeng, Z., and Jacobsen, K. W. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27): 273002, 2017. URL <http://stacks.iop.org/0953-8984/29/i=27/a=273002>.
- Ling, G. Z. S., Foo, J. J., Tan, X.-Q., and Ong, W.-J. Transition into net-zero carbon community from fossil fuels: Life cycle assessment of light-driven co2 conversion to methanol using graphitic carbon nitride. *ACS Sustainable Chemistry & Engineering*, 11(14):5547–5558, 2023.
- Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., and Anandkumar, A. Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789*, 2022.
- Liu, S., Wang, J., Yang, Y., Wang, C., Liu, L., Guo, H., and Xiao, C. Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090*, 2023a.
- Liu, S., Zhu, Y., Lu, J., Xu, Z., Nie, W., Gitter, A., Xiao, C., Tang, J., Guo, H., and Anandkumar, A. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*, 2023b.
- Liu, Z., Zhang, W., Xia, Y., Wu, L., Xie, S., Qin, T., Zhang, M., and Liu, T.-Y. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688*, 2023c.

- Mondal, U. and Yadav, G. D. Methanol economy and net zero emissions: critical analysis of catalytic processes, reactors and technologies. *Green Chemistry*, 23(21):8361–8405, 2021.
- Mukhtar, A., Saqib, S., Lin, H., Shah, M. U. H., Ullah, S., Younas, M., Rezakazemi, M., Ibrahim, M., Mahmood, A., Asif, S., et al. Current status and challenges in the heterogeneous catalysis for biodiesel production. *Renewable and Sustainable Energy Reviews*, 157:112012, 2022.
- Nørskov, J. K., Abild-Pedersen, F., Studt, F., and Bligaard, T. Density functional theory in surface chemistry and catalysis. *Proceedings of the National Academy of Sciences*, 108(3):937–943, 2011.
- NVIDIA Corporation. Megamolbart v0.2, 2022. URL [https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/megamolbart\\_0\\_2](https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/megamolbart_0_2).
- OpenAI. Gpt-4 technical report, 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Pakhare, D. and Spivey, J. A review of dry (co<sub>2</sub>) reforming of methane over noble metal catalysts. *Chem. Soc. Rev.*, 43:7813–7837, 2014. doi: 10.1039/C3CS60395D. URL <http://dx.doi.org/10.1039/C3CS60395D>.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32, pp. 8024–8035. Curran Associates, Inc., 2019. URL <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- Perdew, J. P., Burke, K., and Wang, Y. Generalized gradient approximation for the exchange-correlation hole of a many-electron system. *Physical review B*, 54(23):16533, 1996.
- Rubin, S. M. and Reddy, R. The locus model of search and its use in image interpretation. *Cambridge, Massachusetts*, pp. 590–595, 1977.
- Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.
- Schwaller, P., Probst, D., Vaucher, A. C., Nair, V. H., Kreutter, D., Laino, T., and Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152, 2021.
- Seidl, P., Vall, A., Hochreiter, S., and Klambauer, G. Enhancing activity prediction models in drug discovery with the ability to understand human language. *arXiv preprint arXiv:2303.03363*, 2023.
- Sprueill, H. W., Edwards, C., Olarte, M. V., Sanyal, U., Ji, H., and Choudhury, S. Monte carlo thought search: Large language model querying for complex scientific reasoning in catalyst design. *arXiv preprint arXiv:2310.14420*, 2023.
- Stolarczyk, J. K., Bhattacharyya, S., Polavarapu, L., and Feldmann, J. Challenges and prospects in solar water splitting and co<sub>2</sub> reduction with inorganic and hybrid nanostructures. *ACS Catalysis*, 8(4):3602–3635, 2018.
- Su, B., Du, D., Yang, Z., Zhou, Y., Li, J., Rao, A., Sun, H., Lu, Z., and Wen, J.-R. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022.
- Tawalbeh, M., Javed, R. M. N., Al-Othman, A., Almomani, F., and Ajith, S. Unlocking the potential of CO<sub>2</sub> hydrogenation into valuable products using noble metal catalysts: A comprehensive review. *Environmental Technology & Innovation*, 31:103217, 2023. ISSN 2352-1864. doi: <https://doi.org/10.1016/j.eti.2023.103217>. URL <https://www.sciencedirect.com/science/article/pii/S2352186423002134>.
- Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.
- Tisseraud, C., Comminges, C., Belin, T., Ahouari, H., Soualah, A., Pouilloux, Y., and Le Valant, A. The Cu-Zn synergy in methanol synthesis from CO<sub>2</sub>, part 2: Origin of the methanol and co selectivities explained by experimental studies and a sphere contact quantification model in randomly packed binary mixtures on Cu-Zn coprecipitate catalysts. *Journal of Catalysis*, 330:533–544, 2015. ISSN 0021-9517. doi: <https://doi.org/10.1016/j.jcat.2015.04.035>. URL <https://www.sciencedirect.com/science/article/pii/S0021951715001396>.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

- Tysinger, E. P., Rai, B. K., and Sinitskiy, A. V. Can we quickly learn to “translate” bioactive molecules with transformer models? *Journal of Chemical Information and Modeling*, 63(6):1734–1744, 2023.
- Unsleber, J. P., Liu, H., Talirz, L., Weymuth, T., Mörchen, M., Grofe, A., Wecker, D., Stein, C. J., Panyala, A., Peng, B., et al. High-throughput ab initio reaction mechanism exploration in the cloud with automated multi-reference validation. *The Journal of Chemical Physics*, 158(8), 2023.
- Vall, A., Hochreiter, S., and Klambauer, G. Bioassayclr: Prediction of biological activity for novel bioassays based on rich textual descriptions. In *ELLIS ML4Molecules workshop*, 2021.
- Vaucher, A. C., Schwaller, P., Geluykens, J., Nair, V. H., Iuliano, A., and Laino, T. Inferring experimental procedures from text-based representations of chemical reactions. *Nature communications*, 12(1):2573, 2021.
- Wang, Y., Cao, L., Libretto, N. J., Li, X., Li, C., Wan, Y., He, C., Lee, J., Gregg, J., Zong, H., Su, D., Miller, J. T., Mueller, T., and Wang, C. Ensemble effect in bimetallic electrocatalysts for co2 reduction. *Journal of the American Chemical Society*, 141(42):16635–16642, 2019. doi: 10.1021/jacs.9b05766. URL <https://doi.org/10.1021/jacs.9b05766>. PMID: 31509393.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.
- Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.
- Weininger, D., Weininger, A., and Weininger, J. L. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101, 1989.
- Xi, J., Wang, Z., Wang, W., and Lu, G. In-situ xps study for reaction mechanism of methanol decomposition over cu-ni/zn catalyst. *ACTA PHYSICOCHEMICA SINICA*, 18(1):82–86, 2002.
- Xu, H. and Wang, S. Protranslator: zero-shot protein function prediction using textual description. In *Research in Computational Molecular Biology: 26th Annual International Conference, RECOMB 2022, San Diego, CA, USA, May 22–25, 2022, Proceedings*, pp. 279–294. Springer, 2022.
- Xu, M., Yuan, X., Miret, S., and Tang, J. Protst: Multimodality learning of protein sequences and biomedical texts. *arXiv preprint arXiv:2301.12040*, 2023.
- Xu, S. and Carter, E. A. Theoretical insights into heterogeneous (photo) electrochemical co2 reduction. *Chemical reviews*, 119(11):6631–6669, 2018.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.
- Ye, G., Cai, X., Lai, H., Wang, X., Huang, J., Wang, L., Liu, W., and Zeng, X. Drugassist: A large language model for molecule optimization. *arXiv preprint arXiv:2401.10334*, 2023.
- Zeng, Z., Yao, Y., Liu, Z., and Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022.
- Zhang, X., Wang, L., Helwig, J., Luo, Y., Fu, C., Xie, Y., Liu, M., Lin, Y., Xu, Z., Yan, K., et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023.
- Zhao, H., Liu, S., Ma, C., Xu, H., Fu, J., Deng, Z.-H., Kong, L., and Liu, Q. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *bioRxiv*, pp. 2023–05, 2023a.
- Zhao, W., Zhou, D., Cao, B., Zhang, K., and Chen, J. Adversarial modality alignment network for cross-modal molecule retrieval. *IEEE Transactions on Artificial Intelligence*, 2023b.
- Zhao, Z., Ma, D., Chen, L., Sun, L., Li, Z., Xu, H., Zhu, Z., Zhu, S., Fan, S., Shen, G., et al. Chemdfm: Dialogue foundation model for chemistry. *arXiv preprint arXiv:2401.14818*, 2024.
- Zitnick, C. L., Chanussot, L., Das, A., Goyal, S., Heras-Domingo, J., Ho, C., Hu, W., Lavril, T., Palizhati, A., Riviere, M., et al. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv:2010.09435*, 2020.

## A. GNN Driven Adsorption Energy Calculation

Here, we detail our methodology for GNN-driven adsorption energy calculations. Given a linguistic catalyst recommendation from the LLM, an adsorption energy is produced in a three step process, catalyst generation, adsorbate+catalyst sampling, relaxation.

First, the linguistic representation of a catalyst from the language model must be translated into a formate which is computationally digestible. For instance, a simple catalyst, like pure Copper Zinc (CuZn), can be expressed in several different ways:

- Copper-Zinc
- Copper alloyed with Zinc
- CuZn
- Copper/Zinc-oxide (catalyst not supported by our method)
- Cu-Zn.

Attempting to capture all possible disambiguation of catalyst is intractable for large tree searches, especially with less-well behaved models such as GPT-3.5-turbo. It also becomes more difficult to decide which catalysts to penalize as containing non-metals, such as Copper/Zinc-oxide. Therefore, we prompt the LLM to parse out the correct chemical symbols into digestible lists of chemical symbols (see Appendix B.5 for the prompt details). We would expect the above phrases to be parsed into:

- ["Cu", "Zn"]
- ["Cu", "Zn"]
- ["Cu", "Zn"]
- ["Cu", "Zn", "O"] (catalyst not supported by our method)
- ["Cu", "Zn"].

These examples are much easier to parse out, since we expect to see only chemical symbols. Additionally, the "O" can easily be parsed out to return a penalty value. With these lists of chemical symbols we can use a set of rules to generate the associated 3D structures.

In the second step, the 3D structure of the catalyst is inferred from known reference structures, as specified by the Atomic Simulation Environment (ASE) package (Larsen et al., 2017). For the purposes of this paper, we restrict our focus only to catalyst whose reference structures are in

face-centered cubic (FCC), body-centered cubic (BCC), or hexagonal close-packed (HCP) lattice structures. Catalysts with other reference structures are skipped and assigned a negative reward value. However, many important metals for catalysis have these three lattice structures. If the recommended catalyst is a bimetallic or trimetallic catalyst (composed of two or three elements), we randomly introduce atoms from the secondary and tertiary elements into the structure. For bimetallic compounds, AB, we assume a 2:1 ratio of A to B, while trimetallic catalysts have a 1:1:1 ratio. Catalysts with more than 3 elements return a penalty value. The lattice structure is always inferred by the first species listed by the LLM. For instance CuZn would have the structure of copper, with some atoms randomly replaced with Zn in a 2:1 ratio. To promote the stability of the catalysts, we sample 16 possible configurations and chose only the structure with the lowest energy, as determined by the GNN, to move on to the rest of the GNN calculation.

The third step of the process is the sampling of adsorbate+catalyst configurations. the goal is to sample different locations and orientations for an adsorbate to bind to the surface of the catalyst, to ensure optimal binding is observed. We determine binding sites, use the placement sites specified by ASE for each of the three lattice configuration. Then, for an adsorbate, which we denote generically as, \*XYZ, we place the atom marked by the \* at the binding location at a height 1.87 Å above the surface. Then, to sample orientations of the adsorbate, we rotate the molecule a random angle up to 15 degrees in the x-axis (axis parallel to the surface), then a random angle up to 360 degrees around the z axis (axis perpendicular to the surface). Finally, since some adsorbate molecules may bind differently to certain elements in multimetallic catalysts, we randomly replace the binding site of the catalyst with each of the available metallic elements, with equal probability. This gives fair change of the adsorbate binding to any of the given catalyst species. We take 16 initial adsorbate+catalyst configurations and relax each of them with the GNN to compute adsorption energies.

The fourth and final step is the the GNN driven relaxation. Given the set of atomic coordinates and atomic numbers of a structure, the GemNet-dT model (Gasteiger et al., 2021) returns an approximation to that structure’s adsorption energy and forces,  $E_{\text{ads}}$  and  $\mathbf{F}_{\text{max}}$ . Starting with the sampled adsorbate+catalyst configurations, the structures are relaxed in-batches of 40, using the L-BFGS from Pytorch (Chanusot\* et al., 2021; Paszke et al., 2019). We stop the relaxation terminates when after 64 stems or when the magnitude of the maximum single atom force is below the early stopping threshold,  $\mathbf{F}_{\text{max}} < 0.05 \text{ eV/Å}$ . Then, the minimum adsorption energy of the 16 samples is returned as the final adsorption energy for the adsorbate+catalyst configuration.

These four steps constitute the computational backbone of the reward calculation. For reaction based queries, the reward function is derived from these energies according to Equation 1.

## B. Dataset Design

We propose two task datasets related to catalyst design: the first is derived from the Open Catalyst (OC) Project (Zitnick et al., 2020) and the second consists of complex reasoning queries designed by catalysis experts. Our multi-disciplinary team involves researchers who actively work on designing new catalysts for bio-fuels development.

### B.1. Prompt Design

To apply CHEMREASONER to catalyst discovery, we define use a prompt template which incorporates and overall query for the LLM to answer and we use the actions defined in 5.1 to modify the fields of the templates to reason about the queries. Here, we discuss the s general prompt template, which is used for CHEMREASONER, CoT, and self-consistency, before diving into the construction of queries within the datasets.

Given a query, we use a prompt template which allows the systematic adjustment of an LLM’s context for answering the query, essentially creating a reasoning process. The reasoning has four changeable variables, the “catalyst type”, “inclusion criteria”, “exclusion criteria”, and “candidate\_list\_statement”. Here, we describe each of these variables in detail.

First, the “catalyst type” is the broad category of catalysts to suggest (i.e. metallic catalysts, bimetallic catalysts, transition metal catalysts). Second, we provide a list of ‘inclusion criteria’ and “exclusion criteria” to provide additional context for which properties are more/less useful for suggesting a catalyst that answers the original query. Such examples could be “include catalysts with low cost, high conversion, high Lewis-acidity” or “exclude catalysts which degrade quickly”. While the variables provide context to the LLM, in CHEMREASONER, we additionally include, in the prompt, the suggested catalysts from the previous answer as additional context. This motivates final variable, “candidate list type,” where the prompt LLM is instructed how to use the previous candidates in its evaluation. For instance, it can produce candidates that are “similar to..”, “different from...”, or “have stronger affinity for oxygen than...” the previous candidates.

The overall prompt template is as follows:

```
{query}
{include_statement}{exclude_statement}
Provide scientific explanations
```

```
for each of the {catalyst_label}.
Finally, return a python list
named final_answer which contains
the top-5 {catalyst_label}.
{candidate_list_statement}
```

Take a deep breath and let’s think step-by-step. Remember, you need to return a python list named final\_answer!

The query in the prompt template is determined by our three datasets: OpenCatalyst, BioFuels,

### B.2. OpenCatalyst Dataset

The Open Catalyst dataset (Zitnick et al., 2020) is an extensive collection of (DFT) calculations for various adsorbate+catalyst, intended for training surrogate models for computational chemistry simulations related to catalysis. Here, we utilize this dataset to benchmark the ability of the CHEMREASONER to produce catalyst recommendations for the set of adsorbates given in the Open Catalyst 2020 dataset.

Each query consists of a single adsorbate from the Open Catalyst Project dataset (i.e. \*OHCH<sub>3</sub>, where \* denotes the catalyst binding site). So, each prompt template takes the form Generate a list of top-5 {catalyst\_label} for the adsorption of {adsorbate}, where catalyst\_label begins with ‘metallic catalysts’.

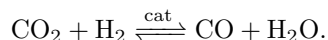
### B.3. BioFuels Dataset

The BioFuels dataset targets more complex adsorbates related to (a) the production of sustainable fuels from organic materials and (b), the reverse water gas shift reaction. The reactions and adsorbates targeted here are crucial to the production synthetic bio fuels with greater selectivity.(Canakci & Van Gerpen, 1999; Daza & Kuhn, 2016; Kattel et al., 2017; Artz et al., 2018; Stolarczyk et al., 2018; Xu & Carter, 2018; Mukhtar et al., 2022)

For the queries in (a), we benchmark the ability of CHEMREASONER to recommend catalysts for the adsorption of specific adsorbates related to bio-fuel production. Such adsorbates are phenol(-ate), anisole, furfural, methanol, methyl, ethanol, acetic acid, and acetate. Additionally, we apply a criterion of interest to CHEMREASONER, such as high/low selectivity, conversion, binding energy. Finally, we provide additional context by specifying a type of reaction we are interested in, hydrodeoxygenation or hydrogenation. The template for these queries is Generate a list of top-5 {catalyst\_label} can bind {adsorbate} in the {reaction\_type} reaction with {additional\_constraint},

where `catalyst_label` begins with `'metallic catalysts'` while `reaction_type`, `adsorbate`, and `additional_constraint` take the values described above.

For the queries in (b), we benchmark the ability of CHEMREASONER to recommend catalysts for the production of CO through the RWGS reaction,



There are several query variations for this reaction. First, we ask CHEMREASONER to recommend catalysts that strongly bind the reactants ( $\text{CO}_2$  and  $\text{H}_2$ ) and weakly bind the product (CO). The remainder of the questions relate to catalysts which demonstrate high convergence for the reaction. Then, we change the type of catalyst requested (metallic catalyst, bimetallic catalyst, bimetallic catalyst with noble metal and base metal). Finally, we optionally include an additional criterion for low cost catalysts. The final query template is `Generate a list of top-5 {catalyst_label} {additional_criterion} for {reaction_question}`.

#### B.4. CO<sub>2</sub>-Conversion Dataset

For the final set of queries, the CO<sub>2</sub>-conversion dataset, we propose a set of 20 queries to benchmark CHEMREASONER’s ability to recommend catalysts for more complex reactions, the conversion of CO<sub>2</sub> to methanol and ethanol (X-anol). Compared to the RWGS reaction, these reactions consist of pathways with different chemical intermediates. The language model will have to consider how the structure of the catalyst promotes the progression of the reaction of several chemical intermediates. Thus, queries involving the CO<sub>2</sub> to X-anol conversion reactions use our reaction pathway based reward.

These query templates echo the same structure as the RWGS templates, `Generate a list of top-5 {catalyst_label} {additional_criterion} for {reaction_question}`.

#### B.5. Symbols Parsing Prompt

One critical requirement of CHEMREASONER’s pipeline is the ability to a fuzzy, linguistic catalyst name into a list of symbols (see Appendix A). To accomplish this task, we use the following LLM prompt template: Consider the following list of catalysts:

`[{cat_0}, {cat_1}, ..., {cat_k}]`.

For each catalyst, return the list of chemical symbols that make up the catalyst. If a catalyst does not have a chemical symbol, return `None`. If a

catalyst is already a chemical formula, repeat the elements in the chemical formula.

Format your list as:

```
{cat_0}: [list_0]
{cat_1}: [list_1]
...
{cat_k}: [list_k]
```

Where `cat_i` is the *i*th catalyst proposed by the LLM. The symbols lists can be parsed out from the answer by following the provided format in the prompt.

#### B.6. Planner Prompt

Here, we detail how CHEMREASONER determines the next set of catalyst descriptors by implementing a planner prompt. The prompt incorporates context from previous catalyst answers and the current search state to update the search states in a reasonable way. The search state corresponds to the 4 action types listed in section 5.1:

1. **Catalyst Type:** the current catalyst type,
2. **Exclusion Criteria:** the list of catalyst design criteria to exclude,
3. **Inclusion Criteria:** The set of catalyst criteria to include,
4. **Relationship to previous candidate set:** How the new catalyst should relate to the previous catalysts (e.g. similar to, different from).

CHEMREASONER uses actions to modify these four state variables. Whereas CHEMREASONER-Expert uses a hand-defined set of actions designed by catalysis experts, CHEMREASONER-Planner uses the LLM to determine the actions. To this end, we propose the following contextualized prompt template for our planner module, presented in figure 6.

In the planner template, the phrases denoted by `$` are treated as “variables” in the LLM’s context window. They hold information about the following information:

1. `search_state`: a dictionary mapping each CHEMREASONER state variable name to its current value,
2. `action_space`: a description of the possible actions,
3. `root_question`: the root question for CHEMREASONER to answer. For example, “What is a good catalyst for X reaction?”
4. `current_prompt`: the previous catalyst prompt that produced the `current_answer`. It should be a modification of `root_question` with the previous actions chosen by CHEMREASONER,
5. `current_answer`: the current catalyst recommen-



Top-5 Catalysts	GPT-4	CHEMREASONER-Expert	CHEMREASONER-Planner
1	Pd	Cu-Zn	Pd-Au
2	Cu	Fe	Pt-Ru
3	Ru	Ni	Ru-Au
4	Rh	Co	Rh-Pd
5	Pt	Cu-Cr	Pt-Au

Table 2. Final catalyst recommendations from CHEMREASONER and GPT-4 (root of the tree) for the  $\text{CO}_2 \rightarrow$  methanol conversion reactions.

dations and their scientific explanations, as returned by the LLM.

```

$search state = {current state}

$action space = [
  "change the type of catalyst to search for",
  "add a new inclusion criteria",
  "add a new exclusion criteria",
  "change the relationship to the candidate list",
]

$root question = {root prompt}

$current prompt = {previous prompt}

$current answer = {previous answer}

Consider the $search state, $root prompt, $current question
and $current answer.

Your task is to suggest possible actions that could achieve
the intent of the $root prompt.

Your answers should use the following guidelines:

Your suggestions should use scientific explanations from the
answers and explanations in $current answer.

Your suggestions should not include MOFs, Zeolites, non-
metals.

Your suggestions should not repeat categories from $search
state.

Let’s think step-by-step, explain your thought process, with
scientific justifications, then return your answer as a
dictionary
mapping from ["catalyst type", "exclusion criteria",
"inclusion criteria", "relationship to previous candidate
list"] to lists of suggestions

```

Figure 6. The planner prompt template. The LLM is prompted to suggest actions which modify the four state variables. Context about the current search state is provided to the LLM by using “variables,” denoted with a \$ symbol.

## C. Search Analysis

### C.1. Parameter selection

Given the compute intensive nature of each CHEMREASONER query, we carried out hyperparameter sweeps to determine the optimal number of actions ( $N_{actions}$ ), as well as the maximum number of relaxation steps ( $N_{relax}$ ). Figure 9 shows how increasing  $N_{actions}$  impacts the distribution of rewards, and we set it to a default value of 8 for all experiments. Figure 7 shows convergence of the GNN for finding the optimal atomistic structure configuration over 300 iterations for 40 random

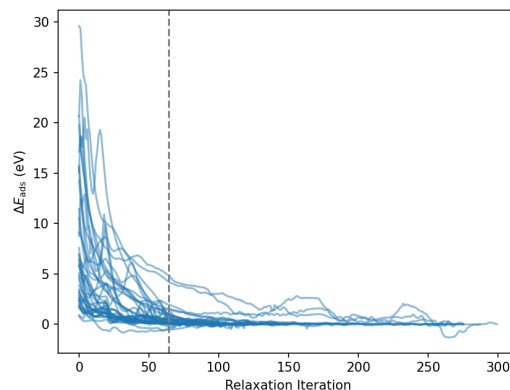


Figure 7. GNN energy versus relaxation iteration for 40 randomly chosen adsorbate+catalyst configurations. The difference between the current and final GNN energies are displayed to highlight the time-to-convergence for each relaxation. The majority of structures are nearly fully relaxed by 64 iterations.

initial adsorbate+catalyst configurations. Relaxations are terminated early if  $|\mathbf{F}_{\max}| < 0.05 \text{ eV}/\text{\AA}$ . At each step in the relaxation process, the difference between the current adsorption energy and the final adsorption energy is reported. The dashed gray line depicts the iterations cutoff for reward calculations in CHEMREASONER, 64 steps. Most iterations are nearly converged by 64 steps.

### C.2. Quantitative Validation with DFT

To validate the adsorption energies being predicted by the GNN, density functional theory (DFT) calculations were performed for the top scoring candidates. These calculations were carried out using the open-source *Quantum ESPRESSO* package (Giannozzi et al., 2009; 2017) which implements DFT using a plane-wave basis set. Core electrons were replaced by pseudopotentials of projector augmented-wave (PAW) type from the *pslibrary* (Dal Corso, 2014). The *PBE* exchange-correlation functional was used (Perdew et al., 1996), and van der Waals effects were included using the DFT-D3 approach (Grimme et al., 2010). All of the calculations were managed and orchestrated using the AiiDA automation framework (Huber

Surface	Pathway	Adsorbate	Adsorption Energy (eV)	
			GNN	DFT
CuZn	Methanol	CO <sub>2</sub>	0.384	-0.066
		CHOH	0.552	5.951
		OCHO	0.577	5.836
		OHCH <sub>3</sub>	-1.160	2.699
CuAlZn	Methanol	CO <sub>2</sub>	0.265	6.816
		CHOH	0.609	-1.824
		OCHO	-0.125	2.820
		OHCH <sub>3</sub>	-1.589	-5.615

Table 3. Comparison of adsorption energies predicted by the GNN to those computed with DFT, for the methanol pathway.

et al., 2020).

While the GNN predicts the adsorption energy directly, each DFT computation of the adsorption energy requires three calculations of the total energy:

$$E_{adsorption}^{DFT} = E_{system}^{DFT} - (E_{slab}^{DFT} + E_{adsorbate}^{DFT}) \quad (2)$$

where  $E_{system}^{DFT}$  is the DFT total energy for the complete system,  $E_{slab}^{DFT}$  is the DFT total energy for the bare slab without any adsorbate, and  $E_{adsorbate}^{DFT}$  total DFT energy of the adsorbate molecule alone.

To compute each of these energies, relaxations were initiated from the final relaxed coordinates coming from the GNN, removing the adsorbate or the slab if necessary. For the both the GNN relaxations and the DFT relaxations, the same atoms in the bulk layers were fixed, while those in the surface layer were allowed to relax. The cell vectors from the GNN structures already included sufficient vacuum above the slab and so were used directly for the corresponding DFT calculations.

### C.3. Analysis of Search Depth

Beyond analyzing the final reward, it is also valuable to understand which of the CHEMREASONERsearch methods discover their best recommended catalysts with fewer computations. One aspect of this search is the depth at which the maximum reward node is found. As shown in Table 4, GPT-4 based searches find their maximum reward node in fewer nodes, on average than, GPT-3.5-turbo based searches, in each dataset. Then, CHEMREASONER-Expert shows superior performance in the OpenCatalyst and BioFuels datasets.

### C.4. Ablation on Actions Taken

To further investigate CHEMREASONER’s performance, we perform an ablation over the number of actions sug-

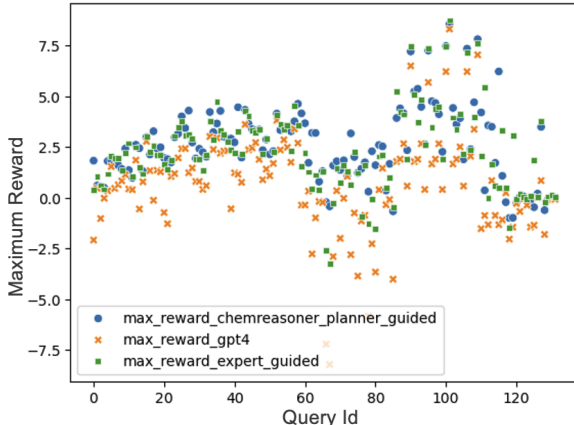


Figure 8. Distribution of simulation rewards for GPT-4 vs the two CHEMREASONER strategies. CHEMREASONER consistently recommends better catalyst candidates across different simulations.

gested from the planning prompt. While it may be unpredictable how many actions the LLM may propose in its plan, we can limit the number of actions, by randomly down-selecting actions, determining the ideal number of actions for CHEMREASONER-Planner. In Figure 9, we show box-and-whisker plots of the best rewards for CHEMREASONER-Planner evaluated over our datasets for various maximum numbers of actions (10 was used in the main text). The final rewards are steadily increasing for max number of actions 2-6, while 8 and 10 offer the highest rewards overall.

### C.5. Scaling of LLM and GNN inference

Both GPT-4 and GPT-3.5-turbo were run using the Python OpenAI Azure interface. Our runtime was dominated by per-deployment rates limits, which are 40K tokens/minute and 240 requests/minute for GPT-4 and 300k tokens/minute and 1800 requests/minute for GPT-3.5-turbo. Prompts at each level in the heuristic search tree are executed asynchronously and in parallel.

Our implementation of the GNN reward function can perform an L-BFGS step in approximately 0.5s with a batch size of 40 adsorbate+catalyst configurations on a single GPU on a DGX2/V100 and on a single A100 40GB.

## D. Multi-modal models for Chemistry

Recent advances in NLP have produced strong results in the chemistry domain by training LLMs (Fabian et al., 2020; Chithrananda et al., 2020; Vaucher et al., 2021; Schwaller et al., 2021; NVIDIA Corporation, 2022; Tysinger et al., 2023) on molecular sequence representations (Weininger, 1988; Weininger et al., 1989; Krenn et al., 2020; Cheng et al., 2023). Beyond this, multi-modal text-molecule models (Edwards et al., 2021; Vall et al., 2021; Zeng et al., 2022;

Table 4. Comparison of the average search depth of best node in each of the search trees.

Method	Category	GPT-3.5-turbo	GPT-4	Search depth reduction (%)
CHEMREASONER-Expert	OpenCatalyst	3.87	3.45	10.97
CHEMREASONER-Expert	BioFuels	3.71	3.50	5.76
CHEMREASONER-Expert	CO <sub>2</sub> -Fuel	4.30	3.30	23.25
CHEMREASONER-Planner	OpenCatalyst	3.79	3.67	3.08
CHEMREASONER-Planner	BioFuels	3.5	3.21	8.16
CHEMREASONER-Planner	CO <sub>2</sub> -Fuel	4.25	3.55	16.47



Figure 9. Impact of maximum number of actions on candidate search. Maximum actions if the most actions (e.g., “exclude low selectivity”) that the LLM is allowed to consider at once when it is reasoning over catalysts.

Xu & Wang, 2022; Su et al., 2022; Seidl et al., 2023; Xu et al., 2023; Zhao et al., 2023b; Liu et al., 2023b) have been proposed to enable high-level control over molecular design (Zhang et al., 2023). This line of work largely focuses on cross-modal retrieval (Edwards et al., 2021; Zeng et al., 2022), translation (Edwards et al., 2022b; Liu et al., 2023c; Christofidellis et al., 2023), and editing (Liu et al., 2022). Building on this trend, recent work has also emerged to enable this behavior via instruction-following (Zhao et al., 2023a; Fang et al., 2023; Guo et al., 2023) and dialogue-based systems for molecules (Liu et al., 2023a; Cao et al., 2023; Ye et al., 2023; Zhao et al., 2024).

## E. Example of Automated Planner-driven Execution

In this example, we demonstrate the reasoning pathway followed by CHEMREASONER-Planner from the root node in a search tree to the best answer.

We provide the question and then present the internal reasoning generated by the LLM to determine the next search

criteria. We interleave each planning phase with the question and answers.

**Question:** Generate a list of top-5 catalyst-label that exhibit high activity for the CO<sub>2</sub> to methanol conversion reaction. Provide scientific explanations and return a list of top 5 answers and their explanations as a list of pairs. Let’s think step-by-step.

### Level 0: Answer from User-specified Query

1. Copper (Cu): Cu-based catalysts are the most widely used in the CO<sub>2</sub> to methanol conversion process, mainly due to their high catalytic activity and selectivity towards methanol. The reaction mechanism involves the initial dissociation of CO<sub>2</sub> on the Cu surface into CO and O, followed by the hydrogenation of CO to form methanol.
2. Zinc (Zn): Zinc is commonly used in conjunction with copper (Cu/Zn) for the CO<sub>2</sub> to methanol conversion. Zinc does not directly participate in the catalysis but enhances the dispersion of copper particles, which increases the number of active sites and thereby the catalytic activity.
3. Palladium (Pd): Palladium can catalyze the CO<sub>2</sub> to methanol conversion reaction via a formate pathway. The CO<sub>2</sub> molecule adsorbs on the palladium surface and is reduced to formate, which subsequently gets hydrogenated to produce methanol.
4. Ruthenium (Ru): Ruthenium-based catalysts exhibit high activity for the CO<sub>2</sub> to methanol conversion reaction. The CO<sub>2</sub> molecule adsorbs on the Ru surface and undergoes dissociation into CO and O, which are then hydrogenated to form methanol.
5. Nickel (Ni): Nickel-based catalysts can efficiently catalyze the CO<sub>2</sub> to methanol conversion via a CO pathway. The CO<sub>2</sub> molecule initially dissociates into CO and O on the Ni surface, followed by the hydrogenation of CO to form methanol.

### Level 1: Internal monologue of Planner

To address the root question, we need to modify the search state to find metallic catalysts that are effective in the CO<sub>2</sub> to methanol conversion reaction. Given the existing state,

we need to add inclusion criteria that align with this requirement.

Let's consider the following factors:

1) **Catalyst Type:** The root prompt asks for metallic catalysts, so we don't need to change the type of catalyst. However, to narrow down our search, we could specify a subclass of metallic catalysts that are known to be effective in this reaction. For example, transition metal catalysts such as copper, zinc, and palladium are known to be active in CO<sub>2</sub> to methanol conversion.

2) **Inclusion Criteria:** Since we're looking for catalysts that exhibit high activity for the CO<sub>2</sub> to methanol conversion reaction, we should include this as a criteria. We might also want to include catalysts that are stable under reaction conditions and have a high selectivity towards methanol.

3) **Exclusion Criteria:** Since we're looking for metallic catalysts, we can exclude non-metallic catalysts. Additionally, we might want to exclude catalysts that degrade quickly or are toxic.

4) **Relationship to Candidate List:** We're looking for catalysts that are similar to those in our candidate list. This means they should have similar properties and performance in the CO<sub>2</sub> to methanol conversion reaction.

Here are my suggestions:

- **Catalyst Type:** ["transition metal catalysts"] - **Inclusion Criteria:** ["high activity in CO<sub>2</sub> to methanol conversion", "stability under reaction conditions", "high selectivity towards methanol"] - **Exclusion Criteria:** ["non-metallic catalysts", "quickly degrading catalysts", "toxic catalysts"] - **Relationship to Candidate List:** ["similar to"]

This way, we can find catalysts that are not only similar to those in our candidate list, but also meet the specific requirements of the root question.

### Level 1: Answer from Planner-generated Query

Based on the requirement of high activity for the CO<sub>2</sub> to methanol conversion reaction, the following metallic catalysts are recommended:

- Rhodium (Rh):** Rhodium-based catalysts have shown excellent performance in CO<sub>2</sub> hydrogenation to methanol. The high activity of Rh is attributed to its ability to dissociate the C-O bond in CO<sub>2</sub>, which is a key step in the conversion process. Additionally, Rh has an appropriate binding energy for CO<sub>2</sub>, allowing for efficient adsorption and desorption of the reactant and product.
- Platinum (Pt):** Platinum is known for its high catalytic activity in various reactions. In the case of CO<sub>2</sub> to methanol conversion, Pt can activate the CO<sub>2</sub> molecule by

weakening the C-O bond, facilitating the subsequent hydrogenation steps.

3. **Iridium (Ir):** Iridium catalysts are effective in CO<sub>2</sub> hydrogenation to methanol due to their high resistance to CO poisoning, a common issue in these reactions. Furthermore, Ir has a suitable binding energy for CO<sub>2</sub>, which enhances the reaction rate by promoting the adsorption and desorption processes.

4. **Gold (Au):** Gold catalysts have recently been found to be effective in CO<sub>2</sub> to methanol conversion. The high activity of Au is associated with its ability to activate CO<sub>2</sub> by facilitating the breaking of the C-O bond. Also, Au has a high resistance to CO poisoning.

5. **Silver (Ag):** Silver is another promising catalyst for CO<sub>2</sub> to methanol conversion. The strong interaction between Ag and CO<sub>2</sub> ensures efficient activation of the CO<sub>2</sub> molecule, leading to high reaction rates. Moreover, Ag is known for its resistance to sintering, which maintains the catalyst's stability and longevity.

### Level 2: Internal monologue of Planner

Based on the current search state, answer and the root question, it is clear that we are looking for metallic catalysts that exhibit high activity for the CO<sub>2</sub> to methanol conversion reaction. The inclusion criterion is commercial viability, and there are no exclusion criteria. The recommended catalysts should be similar to the ones provided in the candidate list.

From the current answer, we see that the recommended catalysts have properties such as high resistance to CO poisoning, efficient activation of CO<sub>2</sub>, ability to dissociate the C-O bond in CO<sub>2</sub>, suitable binding energy for CO<sub>2</sub>, high resistance to sintering, etc.

Based on these, here are some possible actions that we can take:

- **Catalyst Type:** We can look for other types of metallic catalysts that have not been included in the current list. For instance, we can search for "transition metal catalysts" or "noble metal catalysts", since these categories often exhibit high catalytic activities and have similar properties to the catalysts in the current list.

- **Inclusion Criteria:** We can add new inclusion criteria based on the properties of the recommended catalysts in the current answer. For example, we can include "high resistance to CO poisoning" or "ability to dissociate the C-O bond in CO<sub>2</sub>".

- **Exclusion Criteria:** Since all the recommended catalysts in the current answer seem to have desirable properties, it may not be necessary to add any exclusion criteria at this point. However, if we want to narrow down our search, we

could consider excluding catalysts that have "low resistance to sintering" or "poor binding energy for CO<sub>2</sub>".

- **Relationship to Candidate List:** The relationship to the candidate list can remain as "similar to", as we are looking for catalysts that have similar properties to the ones in the current list.

"catalyst-type": ["transition metal catalysts", "noble metal catalysts"], "inclusion-criteria": ["high resistance to CO poisoning", "ability to dissociate the C-O bond in CO<sub>2</sub>"], "exclusion-criteria": ["low resistance to sintering", "poor binding energy for CO<sub>2</sub>"], "relationship-to-candidate-list": ["similar to"]

### Level 2: Answer from Planner-generated Query

Sure, here is the list of top-5 metallic catalysts that exhibit high activity for the CO<sub>2</sub> to methanol conversion reaction with properties of commercial viability and high resistance to CO poisoning:

1. Copper (Cu): Copper-based catalysts, particularly when combined with zinc oxide and alumina (Cu/ZnO/Al<sub>2</sub>O<sub>3</sub>), are the most widely used in the industrial production of methanol from CO<sub>2</sub>. Copper has a high catalytic activity and selectivity towards methanol production. The resistance to CO poisoning is improved by the presence of ZnO, which can absorb CO<sub>2</sub> and present it to the copper sites in a more reactive form.
2. Palladium (Pd): Palladium catalysts, particularly when supported on carbon (Pd/C), show high activity for CO<sub>2</sub> reduction to methanol. Pd has the ability to dissociate CO<sub>2</sub> into CO and O, facilitating the subsequent hydrogenation of CO to methanol. The carbon support provides resistance to CO poisoning by facilitating the desorption of CO from the Pd surface.
3. Ruthenium (Ru): Ruthenium catalysts, especially when supported on carbon (Ru/C), show good activity and selectivity for CO<sub>2</sub> reduction to methanol. Ru has a high CO dissociation energy, which helps resist CO poisoning. The carbon support helps disperse the Ru particles to maximize the catalytic surface area.
4. Nickel (Ni): Nickel catalysts, particularly when supported on alumina (Ni/Al<sub>2</sub>O<sub>3</sub>), have been shown to catalyze the CO<sub>2</sub> to methanol reaction. Nickel has a lower CO binding energy than other metals, which helps resist CO poisoning. The alumina support helps disperse the Ni particles and stabilize the catalyst.
5. Cobalt (Co): Cobalt catalysts, particularly when supported on ceria (Co/CeO<sub>2</sub>), can catalyze the CO<sub>2</sub> to methanol reaction. Cobalt has a relatively low CO binding energy, which helps resist CO poisoning. The ceria support can enhance the CO<sub>2</sub> adsorption and activation,

improving the catalytic activity.

Comparing to the initial list, these metals are generally cheaper and more abundant, making them more commercially viable. They also show a good resistance to CO poisoning due to their lower CO binding energies and/or the use of specific supports.

### Level 3: Internal monologue of Planner

Based on the \$searchState, \$rootPrompt, \$currentQuestion, and \$currentAnswer, here are some possible action suggestions:

1) "catalyst-type": The current catalyst type is "metallic catalysts". The \$currentAnswer provides a list of metallic catalysts that are commercially viable and have high resistance to CO poisoning. Based on this, we could further specify the type of metallic catalysts to explore. For example, we could look into "transition metal catalysts" or "noble metal catalysts". These categories are subclasses of metallic catalysts and could potentially offer new candidates that have not been considered yet.

2) "inclusion-criteria": The current inclusion criteria are "commercial viability" and "high resistance to CO poisoning". The \$currentAnswer explains that these properties are achieved through the choice of metal and the use of specific supports. This suggests that we could add criteria related to these factors. For instance, we could add "use of specific supports to enhance catalytic activity and resistance to CO poisoning" as a new inclusion criteria.

3) "exclusion-criteria": The current exclusion criteria is an empty list. Based on the \$currentAnswer, all the suggested catalysts are not only commercially viable and resistant to CO poisoning, but also have good catalytic activity for the CO<sub>2</sub> to methanol conversion reaction. Therefore, we could add "low catalytic activity for the CO<sub>2</sub> to methanol conversion reaction" as an exclusion criterion.

4) "relationship-to-candidate-list": The current relationship to the candidate list is "similar to". Since the \$currentAnswer provides a new list of catalysts that are different from the initial list, we could change the relationship to "different from".

"catalyst-type": ["transition metal catalysts", "noble metal catalysts"], "inclusion-criteria": ["use of specific supports to enhance catalytic activity and resistance to CO poisoning"], "exclusion-criteria": ["low catalytic activity for the CO<sub>2</sub> to methanol conversion reaction"], "relationship-to-candidate-list": ["different from"]

These changes would allow us to explore new categories of metallic catalysts while focusing on the important properties and excluding those with low activity for the CO<sub>2</sub> to methanol conversion reaction.

**Level 3: Answer from Planner-generated Query**

Let's begin by analyzing the original list of catalysts:

1. Copper (Cu): Commercially viable and shows high resistance to CO poisoning. It has been extensively used in industry for the conversion of CO<sub>2</sub> to methanol due to its high catalytic activity. Cu-based catalysts, especially when combined with zinc oxide and alumina, exhibit remarkable performance.
2. Palladium (Pd): While Pd-based catalysts show high catalytic activity for many reactions, they are known to be susceptible to CO poisoning, which could limit their effectiveness in the CO<sub>2</sub> to methanol conversion.
3. Ruthenium (Ru): Ru-based catalysts exhibit high catalytic activity and resistance to CO poisoning. However, their high cost and limited availability may hinder their commercial viability.
4. Nickel (Ni): Ni-based catalysts are commercially viable due to their relatively low cost. However, they exhibit moderate activity for the CO<sub>2</sub> to methanol conversion and could be susceptible to CO poisoning.
5. Cobalt (Co): Co-based catalysts are less active for the CO<sub>2</sub> to methanol conversion compared to the other catalysts listed. Also, they are susceptible to CO poisoning.

Based on the above analysis, let's create a new list of metallic catalysts that meet the specified requirements:

1. Rhodium (Rh): Rh-based catalysts show high catalytic activity for the CO<sub>2</sub> to methanol conversion and have high resistance to CO poisoning. However, their high cost may limit their commercial viability.
2. Silver (Ag): Ag-based catalysts exhibit high catalytic activity for the CO<sub>2</sub> to methanol conversion, and are resistant to CO poisoning. They are also commercially viable due to their relative abundance and lower cost compared to noble metals.
3. Iron (Fe): Fe-based catalysts are commercially viable due to their low cost and abundance. They show moderate to high activity for the CO<sub>2</sub> to methanol conversion and have good resistance to CO poisoning.
4. Platinum (Pt): Pt-based catalysts have high catalytic activity and resistance to CO poisoning. However, their high cost may limit their commercial viability.
5. Gold (Au): Au-based catalysts show high catalytic activity for the CO<sub>2</sub> to methanol conversion and have high resistance to CO poisoning. However, their high cost may limit their commercial viability.

This list includes catalysts that exhibit high catalytic activity for the CO<sub>2</sub> to methanol conversion and high resistance to

CO poisoning. However, the commercial viability of some of these catalysts may be limited due to their high cost.

**Level 4: Internal monologue of Planner**

From the \$currentAnswer, it seems that we have already achieved the intent of the \$rootPrompt. We have a list of top-5 metallic catalysts that exhibit high activity for the CO<sub>2</sub> to methanol conversion reaction, and have given scientific explanations for each of the metallic catalysts. The list has also been compared to the initial list of catalysts.

However, if we wanted to refine the search further, we could consider the following actions:

1) "Change the type of catalyst to search for": We could limit our search to only noble metallic catalysts. Noble metals like Rhodium, Platinum, and Gold have already been identified as having high catalytic activity and resistance to CO poisoning. However, their commercial viability may be limited due to their high cost. Including less costly noble metals like Silver might provide a good balance between catalytic performance and commercial viability.

2) "Add a new inclusion criteria": We could add "relative abundance" as a new inclusion criteria. This would ensure that we only consider catalysts that are not only commercially viable and resistant to CO poisoning, but also readily available.

3) "Add a new exclusion criteria": We could add "high cost" as a new exclusion criteria. This would help us exclude catalysts that are not commercially viable, despite their high catalytic activity and resistance to CO poisoning.

4) "Change the relationship to the candidate list": Given that we already have a list of suitable catalysts, we could modify the relationship to be "different from". This would help us identify new and potentially better catalysts.

Here's how we could represent these suggested actions:

"catalyst-type": ["noble metallic catalysts"], "inclusion-criteria": ["relative abundance"], "exclusion-criteria": ["high cost"], "relationship-to-candidate-list": ["different from"]

These suggestions would help us refine our search and potentially identify new metallic catalysts that are commercially viable, resistant to CO poisoning, readily available, and not excessively costly.

**Level 4: Answer from Planner-generated Query**

The list provided contains both noble (Rh, Ag, Pt, Au) and non-noble (Fe) metallic catalysts. For the CO<sub>2</sub> to methanol conversion, a noble metallic catalyst is desirable for its high resistance to CO poisoning which is important as CO is an intermediate in this reaction. Iron, however, is not a noble

metal and does not have this property.

1. Palladium (Pd): Palladium is a noble metal that can be used for CO<sub>2</sub> to methanol conversion. It has high resistance to CO poisoning. It is also commercially viable since it is widely used in the industry for hydrogenation reactions. Moreover, palladium can dissociate CO<sub>2</sub> into CO and O, which is beneficial for methanol synthesis.
2. Ruthenium (Ru): Ruthenium is another noble metal that can be used for this conversion. It shows high resistance to CO poisoning and is commercially viable. Ruthenium can adsorb CO<sub>2</sub> and hydrogen simultaneously, facilitating the conversion to methanol.
3. Rhodium (Rh): Rhodium is a noble metal with high catalytic activity for the CO<sub>2</sub> to methanol conversion. It is resistant to CO poisoning and is commercially viable. It has an excellent ability to dissociate CO<sub>2</sub> and hydrogen, facilitating the conversion to methanol.
4. Platinum (Pt): Platinum is a noble metal that is highly resistant to CO poisoning. It is commercially viable and is often used in catalysis due to its high reactivity. Platinum can dissociate CO<sub>2</sub> into CO and O, which is beneficial for methanol synthesis.
5. Gold (Au): Gold is a noble metal that is resistant to CO poisoning. It is commercially viable and is known to catalyze the CO<sub>2</sub> to methanol conversion, especially when used as nanoparticles. Gold can adsorb and activate CO<sub>2</sub>, facilitating its conversion to methanol.

Therefore, the updated list of noble metallic catalysts suitable for the CO<sub>2</sub> to methanol conversion, considering the given criteria, is ['Palladium (Pd)', 'Ruthenium (Ru)', 'Rhodium (Rh)', 'Platinum (Pt)', 'Gold (Au)'].

#### Level 5: Internal monologue of Planner

Looking at the \$rootPrompt, the task is to generate a list of top-5 metallic catalysts. The \$currentAnswer provides a list of noble metallic catalysts. However, the \$rootPrompt does not restrict the type of metallic catalysts to noble ones. Therefore, one possible action is to expand the search to include non-noble metallic catalysts.

Furthermore, the \$currentAnswer includes scientific explanations for each catalyst. These explanations can be used to refine the inclusion and exclusion criteria. For example, the ability to dissociate CO<sub>2</sub> into CO and O is a desirable property mentioned multiple times in the explanations. Therefore, this can be added as a new inclusion criterion. On the other hand, low catalytic activity for the CO<sub>2</sub> to methanol conversion reaction is the current exclusion criterion. We can add another exclusion criterion based on the explanations provided. For example, we can exclude catalysts that cannot adsorb and activate CO<sub>2</sub>, as this property

is also mentioned as being beneficial for the conversion to methanol.

The relationship to the candidate list can be changed to "different from" to find new metallic catalysts that are not in the current list.

Here are the suggestions based on the above analysis:

"catalyst-type": ["non-noble metallic catalysts"], "inclusion-criteria": ["ability to dissociate CO<sub>2</sub> into CO and O", "ability to adsorb and activate CO<sub>2</sub>"], "exclusion-criteria": ["cannot adsorb and activate CO<sub>2</sub>"], "relationship-to-candidate-list": ["different from"]

These changes will help in finding new metallic catalysts that are not noble metals but still have high catalytic activity for the CO<sub>2</sub> to methanol conversion reaction. This is in line with the intent of the \$rootPrompt.

#### Level 5: Answer from Planner-generated Query

The CO<sub>2</sub> to methanol conversion reaction is a crucial process in catalysis and has been extensively studied. Here are my top-5 recommendations for noble metallic catalysts that exhibit high activity for this reaction, along with the scientific explanations:

1. **Palladium-Gold (Pd-Au) Alloy**: This alloy catalyst has shown excellent catalytic performance for CO<sub>2</sub> reduction to methanol. The Pd-Au alloy, especially when supported on titanium dioxide (TiO<sub>2</sub>), provides a synergistic effect where Pd facilitates CO<sub>2</sub> activation and dissociation, while Au helps in reducing the CO intermediate to methanol. The alloy structure also enhances the resistance to CO poisoning.
2. **Platinum-Ruthenium (Pt-Ru) Alloy**: Pt-Ru alloy, when supported on carbon (C), exhibits high activity and selectivity for CO<sub>2</sub> to methanol conversion. The bifunctional mechanism involves CO<sub>2</sub> activation and dissociation on Ru sites and subsequent hydrogenation to methanol on Pt sites.
3. **Ruthenium-Gold (Ru-Au) Alloy**: The Ru-Au alloy is another excellent choice for this reaction. Ru helps in dissociating CO<sub>2</sub> into CO and O, while Au aids in the reduction of CO to methanol. The alloy structure also enhances the resistance to CO poisoning.
4. **Rhodium-Palladium (Rh-Pd) Alloy**: The Rh-Pd alloy, when supported on alumina (Al<sub>2</sub>O<sub>3</sub>), exhibits high activity for CO<sub>2</sub> to methanol conversion. Rh facilitates CO<sub>2</sub> dissociation, while Pd helps in reducing the CO intermediate to methanol. The alloy structure enhances the resistance to CO poisoning.
5. **Platinum-Gold (Pt-Au) Alloy**: The Pt-Au alloy, when supported on carbon (C), shows high activity and selectivity for the CO<sub>2</sub> to methanol conversion. The bifunc-

tional mechanism involves CO<sub>2</sub> activation and dissociation on Au sites and subsequent hydrogenation to methanol on Pt sites.