CMI-AI: A Physically Grounded Framework for Autonomous Chemical Mechanism Inference via Neural-Symbolic Tree Search

EXAONE-3.5 (32B), GPT-4o, Gemini 1.5 Pro and Claude Sonnet 4
(Names with versions only, sorted by priority)*

Anonymous†
None
None
None

Abstract

While AI has become essential in accelerating chemical research, general-purpose Large Language Models (LLMs) often suffer from a "grounding gap," prioritizing statistical correlations over physical laws and producing "chemical hallucinations" such as the energetically prohibitive cleavage of C–F bonds ($\approx 485\,\mathrm{kJ/mol}$). To address this, we present Chemical Mechanism Inference AI (CMI-AI), a physically grounded framework based on EXAONE [1] that couples a neural proposal model with a symbolic chemistry core (RDKit [2]) to ensure structural and valence validity. CMI-AI formulates mechanism generation as a constrained tree-search over intermediate molecular states, where each candidate step is executed as an explicit graph edit and validated against deterministic signals including Bond Dissociation Energy (BDE) heuristics, pK_a-consistent proton-transfer logic, and intermediate stability indicators. To accommodate the non-linear branching of organic reactions, we employ a Beam Search Algorithm to maintain and rank multiple competing pathways via a hybrid objective function. Furthermore, we introduce the Masking Verification Step (MVS), a diagnostic methodology that masks key causal anchors—such as leaving-group identity and stabilization arguments—to distinguish authentic chemical understanding from superficial template memorization, using BERT-style masking tests [3]. By integrating online physical constraints with a verification-driven self-correction loop, CMI-AI effectively rectifies systematic pattern-matching errors and provides a highly reliable, interpretable pathway for autonomous chemical discovery.

1 Introduction

The acceleration of chemical research through Artificial Intelligence (AI) has transitioned from a theoretical possibility to an operational necessity. As the chemical space expands exponentially, Large Language Models (LLMs) have emerged as pivotal tools for predicting reaction outcomes, suggesting synthetic routes, and interpreting complex spectroscopic data. By processing vast repositories of chemical literature, these models aim to function as autonomous laboratory assistants, streamlining the "Design-Make-Test-Analyze" (DMTA) cycle. However, as the complexity of molecular systems

*Use footnote for providing further information, for less known open models (webpage, version)

†Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

increases, a critical bottleneck has emerged: the inherent inability of general-purpose LLMs to distinguish between statistical correlation and physical causation.

Despite their linguistic fluency, contemporary LLMs are fundamentally "stochastic parrots" [4] that predict the next token based on learned frequencies within a training corpus. In the domain of organic chemistry, this leads to profound logical fallacies. For instance, when presented with trifluoroacetic anhydride ($TFAA$) in an acylation reaction, a model might erroneously predict the cleavage of a C–F bond simply due to the high frequency of fluorine atoms in the input string. This prediction ignores the fundamental principle of bond dissociation energy (BDE); the C–F bond is one of the strongest covalent bonds in organic chemistry ($\approx 485\,\mathrm{kJ/mol}$), making its cleavage under mild conditions energetically prohibitive compared to the much more labile C–O bond ($\approx 358\,\mathrm{kJ/mol}$). Such "chemical hallucinations" are not merely minor errors; they represent a fundamental lack of grounding in the physical laws that govern the material world.

To bridge this gap, we must move beyond data-centric scaling and toward knowledge-augmented reasoning. Scientific discovery requires more than just predicting an output; it demands a rigorous adherence to thermodynamic and kinetic constraints. In this paper, we introduce an enhanced reasoning framework based on EXAONE [1], specifically designed to integrate explicit physical constraints into the inference process. Our approach anchors the decision-making process in fundamental chemical principles through a *Hybrid Reasoning Core* that combines LLMs with symbolic AI logic via RDKit [2] integration. By utilizing a *Tree Search Mechanism Inference* based on the *Beam Search Algorithm* [9], the model explores the reaction space step-by-step, evaluating intermediate stability and bond energies to find the most physically plausible path.

Furthermore, we propose a novel diagnostic methodology termed the *Masking Verification Step* (MVS). This protocol incorporates a *Self-Correction Feedback Loop* where mechanisms are automatically verified; if the understanding score falls below a defined threshold (e.g., 70%), feedback is injected into the next iteration to drive iterative improvement. By selectively obscuring the causal drivers of a reaction mechanism and requiring the model to re-derive the logic through masking tests, we establish a new standard for model interpretability and reliability.

Specifically, our contributions are threefold:

1. **Logical Rectification:** We demonstrate how EXAONE identifies and corrects the systematic "pattern-matching errors" prevalent in leading general LLMs by employing a verification-driven retry mechanism.
2. **Physical Grounding:** We detail the integration of thermodynamic stability data—including bond energy checks, pK_a logic, and leaving group ability—into the model's reasoning chain, ensuring that predicted mechanisms are energetically feasible.
3. **Verification Protocol:** We introduce the MVS as a robust litmus test, utilizing BERT-style `[MASK]` tests to distinguish between superficial memorization and the authentic application of chemical principles.

By synthesizing physical laws with neural architectures, this work provides a pathway toward highly reliable AI agents capable of partnering with human scientists in the pursuit of complex chemical discovery.

2 Related Work

2.1 LLMs for Chemistry and the Grounding Gap

Large language models (LLMs) are increasingly used in cheminformatics as flexible interfaces that map between natural language, SMILES/IUPAC, and structured reaction outputs (e.g., products or mechanistic steps). Despite strong linguistic fluency, general-purpose LLMs often exhibit a *grounding* gap: they produce statistically plausible but chemically inconsistent proposals that violate basic constraints such as realistic bond cleavage, charge conservation, or acid–base feasibility [4]. In mechanism inference, common failures include ignoring bond dissociation energy (BDE) hierarchies and neglecting pK_a-consistent proton-transfer and leaving-group logic under the stated conditions. These limitations motivate approaches that augment generation with explicit physical and chemical checks so that outputs are not only fluent but also admissible.

2.2 Neural–Symbolic Reasoning, Search, and Verification

To reduce hallucinations, hybrid systems combine LLMs with symbolic chemistry toolkits (e.g., RDKit [2]) that enforce structural validity and compute chemically relevant features. While many pipelines apply symbolic filters post hoc, recent directions integrate such checks within the inference loop to prune unphysical candidates early. Mechanism prediction further benefits from explicit search, since real reactions admit competing branches (e.g., substitution vs. elimination, rearrangement vs. capture) that standard greedy decoding fails to preserve; beam-search-style tree exploration is a practical alternative [9]. Beyond plausibility, verifying that a rationale reflects causal understanding rather than template recall remains challenging, motivating diagnostic protocols that perturb or mask key causal factors and trigger correction when inconsistencies are detected. Our work follows this trajectory by using EXAONE-3.5 [1] as a proposal model while coupling RDKit-based constraints, beam-search mechanism exploration, and a masking-based verification step with a retry loop for iterative self-correction.

3 Problem Definition

We study *autonomous chemical mechanism inference* as the task of generating a physically plausible, step-wise mechanistic explanation conditioned on a reaction query. Concretely, a query x consists of reactants, reagents, and conditions (e.g., solvent, temperature, additives), represented as a mix of structured strings (SMILES/IUPAC) [10] and natural-language context. The desired output is a mechanism m that can be expressed as a sequence of intermediate states and elementary transformations,

$$m = (s_0 \to s_1 \to \cdots \to s_T),$$

where each state s_t encodes a molecular graph (or set of graphs) and each transition corresponds to a chemically meaningful operation (bond formation/cleavage, proton transfer, ionization, rearrangement, etc.).

A practical system must satisfy two requirements simultaneously: (i) *chemical validity* (structures and moves are well-formed) and (ii) *physical plausibility* (the path is consistent with energetics and acid–base logic). This is challenging because the search space of m is combinatorial and because general-purpose LLMs are trained to maximize likelihood rather than physical consistency.

We decompose the problem into three bottlenecks that recurrently block reliable deployment: the *Grounding Gap*, *Branching Complexity*, and *Veridicality Scarcity*.

3.1 The Grounding Gap: Statistical Correlation vs. Physical Law

Failure mode. General-purpose Large Language Models (LLMs) often produce chemically fluent but physically inconsistent steps because generation is driven by token statistics. As a canonical example, the presence of fluorinated motifs may bias the model toward transformations that mention C–F cleavage, even though such steps are typically disfavored under mild conditions due to the large bond dissociation energy (BDE) of C–F bonds ($\approx 485\,\mathrm{kJ/mol}$). In contrast, bonds such as C–O or C–Br are often more labile depending on context, making their cleavage more plausible in many reaction settings.

What must be grounded. For mechanism inference, "grounding" means that generation must be constrained by (at least) the following deterministic signals:

- **Bond energetics:** relative BDEs and local energetic preferences that rule out high-cost bond-breaking events.
- **Acid–base feasibility:** pK_a-consistent proton-transfer logic and base/nucleophile strength.
- **Leaving-group ability:** whether a group can depart under the given conditions and whether charge distribution is chemically reasonable.

Problem statement. How can we constrain a probabilistic neural generator with deterministic physical laws (bond energies, pK_a logic, leaving-group ability) so that the model avoids "chemical hallucinations" *during* inference rather than relying on post-hoc filtering?

3.2 Branching Complexity: Linear Generation vs. Non-linear Search

Failure mode. Organic mechanisms are rarely linear narratives; they are non-linear search processes over a high-dimensional intermediate space. A single event (e.g., ionization to a carbocation) can lead to multiple competing continuations (e.g., substitution vs. elimination, rearrangement vs. capture). Standard LLM decoding (greedy or sampling) approximates a single trajectory in token space and is therefore prone to committing early to a locally likely step that later becomes chemically inconsistent.

Search formulation. Let $\mathcal{M}(x)$ denote the (very large) set of candidate mechanisms consistent with the input. We seek a mechanism that is both likely and physically plausible. A convenient abstraction is a constrained optimization

$$m^* = \arg\max_{m \in \mathcal{M}(x)} \log p_\theta(m \mid x) + \lambda \, \mathrm{Score}_{\mathrm{phys}}(m \mid x) \quad s.t. \quad \mathrm{Valid}(m) = 1,$$

where $\log p_\theta$ captures the neural prior and $\mathrm{Score}_{\mathrm{phys}}$ aggregates online checks (energetics, stability, pK_a feasibility, leaving-group constraints). Because exhaustive search is intractable, the system must approximate this objective with a structured exploration strategy (e.g., tree search with a bounded beam).

Problem statement. How can we move beyond linear text generation toward a tree-search mechanism that maintains multiple competing partial mechanisms, evaluates their stability/optimality, and selects physically plausible branches under a fixed compute budget?

3.3 Veridicality Scarcity: Memorization vs. Authentic Understanding

Failure mode. Even when a model outputs a correct mechanism for a well-known reaction (e.g., Wittig [11]), it is unclear whether the output reflects causal understanding or memorization of frequent templates. This ambiguity becomes critical when transferring to novel substrates that are structurally similar but mechanistically sensitive to subtle changes (e.g., leaving-group identity, solvent polarity, or a shifted pK_a regime). Without a verification signal, a system can appear accurate on canonical examples while failing unpredictably on out-of-distribution queries.

Need for a measurable veridicality signal. We require an operational metric that probes whether the model can re-derive causal logic when surface cues are removed. Let $u(m, x) \in [0, 1]$ denote an *understanding score* computed by diagnostic tests that perturb or mask key causal descriptors in the model's own rationale and measure whether the missing logic can be reconstructed consistently.

Problem statement. How can we implement a verification-driven feedback loop and diagnostic tests—specifically Masking Verification Steps (MVS)—to (i) quantify a model's understanding score u and (ii) trigger iterative self-correction when u falls below a threshold (e.g., 0.7), thereby improving reliability on novel mechanism instances?

4 Proposed Method

4.1 Overall System Architecture

We propose *Chemical Mechanism Inference AI* (CMI-AI), a hybrid autonomous agent that performs mechanistic deduction as a *state-machine-controlled* search over chemically valid intermediates. The system integrates: (i) a neural proposal model (LLM) for hypothesis generation, (ii) a symbolic chemistry core (RDKit-based [2]) for deterministic validity checks and chemical feature extraction, and (iii) a search-and-verification controller (LangGraph-driven [5]) that coordinates expansion, pruning, and self-correction.

Problem definition. Given a reaction query x consisting of (1) reactants $\mathcal{R}$, (2) reagents/catalysts $\mathcal{G}$, (3) conditions $\mathcal{C}$ (solvent, temperature, atmosphere, time, concentration when available), and optionally (4) products $\mathcal{P}$, the goal is to infer a mechanism m as a discrete sequence of chemically meaningful elementary steps that transforms $\mathcal{R}$ into $\mathcal{P}$ while satisfying structural and physicochemical constraints implied by $(\mathcal{G}, \mathcal{C})$.

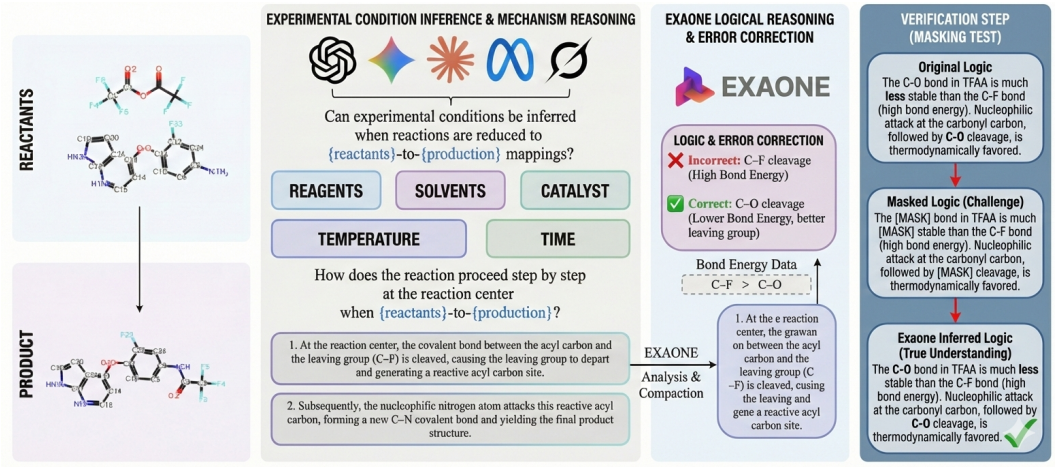

Figure 1: Overview of the proposed CMI-AI pipeline: (i) reactant-to-product mapping and experimental condition inference, (ii) LLM-based mechanism proposal with logical correction under physical constraints, and (iii) masking-based verification (MVS) for veridicality and iterative refinement.

State and action formalism. CMI-AI represents a mechanism as

$$m = (s_0 a_0 s_1 a_1 \cdots a_{T-1} s_T),$$

where each state s_t is a multiset of *molecular graphs*

$$s_t = \{(G_{t,1}, q_{t,1}, \mu_{t,1}), \ldots, (G_{t,n_t}, q_{t,n_t}, \mu_{t,n_t})\}.$$

Here $G_{t,i}$ is an RDKit molecule graph with explicit hydrogens when necessary, $q_{t,i}$ is the formal charge, and $\mu_{t,i}$ is an atom-mapping function from initial atoms to current atoms (used to maintain conservation and to localize reaction centers). States also include a compact *context record* κ_t containing inferred condition flags (e.g., "strong base present", "Lewis acid present", "oxidative conditions", "protic solvent", "radical initiator"), which gate feasible step types.

Each action a_t is an *elementary transformation* drawn from a constrained vocabulary:

$$a_t \in \left\{ \begin{array}{l} bondformation, bondcleavage, heterolysis/homolysis, protontransfer, ionization, \\ addition/elimination, rearrangement, coordination, redox \end{array} \right\}.$$

Operationally, an action is parameterized by: (i) reaction-center atoms (mapped indices), (ii) bond order edits, (iii) charge updates, and (iv) optional auxiliary participants (base/acid/catalyst species from $\mathcal{G}$). We write:

$$s_{t+1} = \mathrm{Apply}(s_t, a_t),$$

where Apply executes RDKit graph edits plus bookkeeping (atom mapping propagation, formal charge reassignment, fragmentation/association).

Preprocessing: mapping and condition inference. Before search, CMI-AI performs (A) reactant-to-product atom mapping [12] to obtain a consistent mapping seed μ_0, and (B) condition inference to populate κ_0. Mapping is used to: (1) identify net bond changes between $\mathcal{R}$ and $\mathcal{P}$, (2) define target reaction centers, and (3) provide a termination signal when product graphs are matched (up to resonance/tautomer equivalence).

Condition inference extracts structured constraints from textual metadata and reagent identities (e.g., "NaOMe/MeOH" $\Rightarrow$ strong base + protic solvent; "hv, AIBN" $\Rightarrow$ radical regime; "Pd(PPh_3)$_4$" $\Rightarrow$ cross-coupling template priors but still validated).

Controller as a state machine. We implement a cyclic controller with states *Propose* $\rightarrow$ *Validate* $\rightarrow$ *Score* $\rightarrow$ *Select* $\rightarrow$ *Verify* $\rightarrow$ (repeat). LangGraph [5] manages the global search state and enforces that a candidate cannot advance unless it passes deterministic checks. This design decouples creative hypothesis generation (LLM) from physical/structural constraints (symbolic core).

4.2 Hybrid Reasoning and Structural Validation

Neural proposal model (LLM). For each partial mechanism prefix $\pi_t = (s_0, a_0, \ldots, s_t)$, the LLM is prompted to propose a bounded candidate set of next actions:

$$\mathcal{A}_t = \{a_t^{(1)}, \ldots, a_t^{(K)}\}, \quad K \leq K_{\max}.$$

To ensure the proposals are executable, the prompt requests a structured output:

$$a_t^{(k)} = (type, centeratoms, bondedits, \Delta q, participants, rationale).$$

We additionally ask the LLM to output a *local justification* explaining why the step is plausible under κ_t (solvent, acid/base strength, leaving group, stabilization, etc.), because this text becomes the basis for the MVS verification stage.

Symbolic structural constraints (RDKit). Each proposed intermediate $s_{t+1}^{(k)} = \mathrm{Apply}(s_t, a_t^{(k)})$ is immediately processed by an RDKit-integrated validation layer:

$$\mathrm{Valid}(s) = 1 \iff (valence, charge, atomtyping, aromaticity, andgraphconsistency) aresatisfied.$$

Concretely, the validator executes the following checks:

- **Valence feasibility:** For every atom, RDKit valence rules are enforced with allowed hypervalence exceptions (e.g., P(V), S(VI)) only when atom type matches.
- **Formal charge bookkeeping:** Total charge change must be explained by explicit ionization/proton transfer actions; "free" charge creation is forbidden.
- **Connectivity constraints:** Disconnected fragments are allowed only if the action implies fragmentation (heterolysis/homolysis) or if a spectator ion is produced with a chemically plausible counterion present.
- **Atom mapping conservation:** The mapping μ must remain bijective for conserved atoms; newly introduced atoms are disallowed except for explicitly provided reagents (e.g., halogenation source) and must be labeled as such.
- **Reaction-center locality:** Each action must involve a limited reaction center (bounded number of bond edits), preventing unrealistic "global rewrites".

Invalid candidates are discarded before scoring. This *validate-before-search* policy prevents unphysical branches from contaminating subsequent exploration.

Logical correction under constraints. If an action nearly passes but fails due to a correctable issue (e.g., missing proton transfer to satisfy charge, incorrect bond order on aromatic system), we invoke a constrained correction step: the LLM is given the validator's failure code and must propose the minimal edit to repair the action while preserving the intended chemistry. The corrected action is re-validated; if it still fails, it is discarded.

4.3 Tree-Search-Based Mechanism Inference

Mechanism inference as constrained optimization. Let $\mathcal{M}(x)$ be the set of all mechanisms consistent with x. We approximate:

$$m^* = \arg\max_{m \in \mathcal{M}(x)} \underbrace{\log p_\theta(m \mid x)}_{LLMprior} + \lambda \underbrace{\mathrm{Score}_{\mathrm{phys}}(m \mid x)}_{physicalplausibility} \quad s.t. \quad \mathrm{Valid}(m) = 1.$$

Since exhaustive enumeration is intractable, we use beam search [9] over partial mechanisms with strict validity gating.

Beam search over partial mechanisms. Let $\mathcal{B}_t$ be the beam at depth t with $|\mathcal{B}_t| \leq B$. Each element is a tuple

$$(\pi_t, S(\pi_t), \eta_t),$$

where π_t is the partial path, $S(\pi_t)$ is its cumulative score, and η_t stores auxiliary traces (e.g., which constraints were binding, extracted chemical features, and explanation text for later MVS).

At each depth:

$$\mathcal{C}_{t+1} = \bigcup_{(\pi_t,\cdot)\in\mathcal{B}_t} \bigcup_{a\in\mathcal{A}(\pi_t)} \{\pi_{t+1} = \pi_t \oplus a\}.$$

We then apply **(i)** deterministic validation, **(ii)** physical scoring, and **(iii)** top-B selection:

$$\mathcal{B}_{t+1} = \mathrm{TopB}(\mathcal{C}_{t+1};\ S(\cdot)).$$

This retains multiple mechanistic hypotheses (e.g., S_N1 vs. S_N2 vs. $E1/E2$, competing rearrangements) until constraints or evidence separates them.

Duplicate and symmetry pruning. To avoid beam collapse into equivalent states, we canonicalize each state s_t using RDKit canonical SMILES with mapping annotations stripped to a canonical representative and maintain a visited hash. Candidates with identical canonical states at the same depth keep only the best-scoring path (or the top-k if diversification is required).

Stopping criteria and completion. A path terminates when one of the following holds: (1) product match: s_t matches $\mathcal{P}$ up to tautomer/resonance equivalence and spectator species removal; (2) dead-end: no valid actions remain; (3) budget: $t = T_{\max}$; (4) stagnation: repeated states detected or score falls below a depth-dependent threshold. We return the highest-scoring complete path and optionally the top-n alternatives with diagnostic traces.

4.4 Physical Grounding and Scoring Functions

Step-wise physical scoring. For each validated transition $s_t a_t s_{t+1}$ we compute:

$$\phi(x, s_t, a_t, s_{t+1}) = w_{\mathrm{bde}}\phi_{\mathrm{bde}} + w_{\mathrm{pk}}\phi_{\mathrm{pk}} + w_{\mathrm{stab}}\phi_{\mathrm{stab}} - w_{\mathrm{pen}}\phi_{\mathrm{pen}},$$

and define the cumulative plausibility:

$$\mathrm{Score}_{\mathrm{phys}}(\pi_t \mid x) = \sum_{i=0}^{t-1} \phi(x, s_i, a_i, s_{i+1}).$$

All components are deterministic functions of (i) RDKit-derived structural features and (ii) reference tables/heuristics indexed by functional group, solvent class, and reagent identity.

Bond dissociation energy (BDE) plausibility. Let $\mathcal{B}_{\mathrm{break}}(a_t)$ and $\mathcal{B}_{\mathrm{form}}(a_t)$ be the sets of bonds broken and formed by a_t. We estimate:

$$\Delta\mathcal{E}_{\mathrm{bde}}(a_t) = \sum_{b\in\mathcal{B}_{\mathrm{break}}(a_t)} \mathrm{BDE}(b) - \sum_{b\in\mathcal{B}_{\mathrm{form}}(a_t)} \mathrm{BDE}(b).$$

We then penalize unusually uphill steps unless the context κ_t indicates strong driving forces (e.g., redox potential, precipitation, gas evolution, strong acid/base):

$$\phi_{\mathrm{bde}} = -\max(0,\ \Delta\mathcal{E}_{\mathrm{bde}}(a_t) - \tau_{\mathrm{bde}}(\kappa_t))\,.$$

For example, breaking a C–F bond without explicit activation is heavily penalized; conversely, formation of a strong bond (e.g., $C=O$ in acyl substitution) offsets the penalty.

pK_a-consistent proton transfer and nucleophilicity logic. For proton transfer steps, we enforce feasibility using inequality constraints derived from pK_a tables. For deprotonation:

$$\mathrm{Base} + \mathrm{H{-}Substrate} \to \mathrm{BaseH}^+ + \mathrm{Substrate}^-,$$

we require:

$$pK_a(\mathrm{BaseH}^+) - pK_a(\mathrm{H{-}Substrate}) \geq \tau_{\mathrm{pk}}(\kappa_t),$$

and set:

$$\phi_{\mathrm{pk}} = \min\big(0,\ pK_a(\mathrm{BaseH}^+) - pK_a(\mathrm{H{-}Substrate}) - \tau_{\mathrm{pk}}\big)\,.$$

For nucleophilic substitution/addition, we compute a leaving-group score and nucleophile score from context-aware heuristics:

$$\phi_{\mathrm{pk}} \leftarrow \phi_{\mathrm{pk}} + \mathrm{LGScore}(\mathit{leavinggroup}, \kappa_t) + \mathrm{NuScore}(\mathit{nucleophile}, \kappa_t),$$

and disallow steps that violate hard rules (e.g., poor leaving group without activation, strongly hindered S_N2 at tertiary centers in protic solvent without special conditions).

Intermediate stability heuristic. We score whether the intermediate class is plausible under conditions. Let $\mathrm{Class}(s)$ be a coarse type (carbocation, radical, anion, zwitterion, metal complex). We compute:

$$\phi_{\mathrm{stab}} = \mathrm{StabilityHeuristic}(x, s_{t+1}),$$

where $\mathrm{StabilityHeuristic}$ uses RDKit-derived descriptors, including: (i) degree of substitution at charged center, (ii) resonance delocalization indicators (allylic/benzylic adjacency, conjugation), (iii) inductive stabilization (EWG/EDG counts within n bonds), (iv) aromaticity preservation/violation, (v) known unstable motifs under κ_t (e.g., free carbocations in strongly basic media). Unstable intermediates (e.g., primary carbocation without resonance in non-superacid conditions) receive a strong negative score or are pruned by hard constraints when appropriate.

Penalty term for constraint violations. ϕ_{pen} aggregates soft penalties that do not strictly invalidate the structure but reduce plausibility: (1) excessive charge separation without counterion stabilization, (2) high ring strain creation without compensation, (3) implausible stereochemical inversion/retention given step type, (4) unnecessary multi-bond edits in a single step.

Combined path score. For a partial path π_t, the controller maintains:

$$S(\pi_t) = \log p_\theta(\pi_t \mid x) + \lambda \, \mathrm{Score}_{\mathrm{phys}}(\pi_t \mid x),$$

where $\log p_\theta$ is accumulated from the LLM's action likelihoods (or calibrated proxy scores when logprobs are unavailable), and λ balances learned prior against deterministic physical plausibility.

4.5 Masking Verification Step (MVS) and Iterative Refinement

Motivation: veridicality vs. template recall. Mechanisms may appear correct while being supported by superficial template matching. We therefore introduce a *Masking Verification Step* (MVS) that explicitly probes whether the model's rationale encodes causal chemical determinants rather than memorized narratives.

Mask generation. Given a candidate mechanism m and its textual rationale r, we extract a set of *causal anchors*: (i) reaction-center atoms (mapped indices), (ii) leaving group identity and its justification, (iii) key pK_a comparisons and acid/base roles, (iv) stabilization arguments (resonance, hyperconjugation, inductive effects), (v) energetic cues (strong/weak bond changes). We then generate M masking queries $\mathcal{Q}(r) = \{q_1, \ldots, q_M\}$ by replacing anchors with `[MASK]` tokens, ensuring each query tests one causal factor at a time. For example:

- leaving group mask: "`[MASK]` leaves because it is stabilized as `[MASK]`".
- pK_a mask: "Deprotonation is favored since $pK_a($`[MASK]`$) - pK_a($`[MASK]`$) \geq$ `[MASK]`."
- reaction-center mask: "Attack occurs at atom `[MASK]` because `[MASK]`."

Reconstruction and symbolic verification. For each q_j, the model reconstructs $\hat{r}_j$. We then run a symbolic verifier $\mathrm{Verify}(\hat{r}_j)$ that checks whether the reconstruction matches the underlying mechanism m and satisfies chemical consistency: (1) reconstructed atoms correspond to actual mapped reaction center, (2) pK_a directionality consistent with the claimed acid/base, (3) leaving group ability consistent with κ_t, (4) stated stabilization matches RDKit-derived features (e.g., benzylic, allylic).

Understanding score. We define:

$$u = \mathrm{UnderstandingScore}(m, r, x) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\{\mathrm{Verify}(\hat{r}_j) = 1\}.$$

Unlike generic self-consistency, u is grounded in deterministic checks tied to the candidate mechanism and extracted structural facts.

Verification-driven retry loop. If $u < u_{\min}$ (e.g., $u_{\min} = 0.7$), the controller diagnoses failure modes and generates targeted feedback f (e.g., "leaving group claim contradicts structure", "proton transfer direction violates pK_a", "reaction center mismatched mapping"). The agent then re-enters the propose/validate cycle with constraints injected:

$$(m^{(i+1)}, r^{(i+1)}) \leftarrow \mathrm{AgentStep}(x, f^{(i)}), \quad f^{(i)} = \mathrm{Diagnose}(m^{(i)}, r^{(i)}, x),$$

until $u \geq u_{\min}$ or a retry budget is exhausted. This makes verification an *active control signal* rather than a passive metric.

Final output. The final mechanism is the highest-scoring complete path that simultaneously satisfies: (i) hard structural validity ($\mathrm{Valid}(m) = 1$), (ii) high physical plausibility (maximizing $\mathrm{Score}_{\mathrm{phys}}$), (iii) veridicality under MVS ($u \geq u_{\min}$). We optionally return a ranked list of alternative mechanisms with their constraint traces and MVS failure reasons to support interpretability and error analysis.

5 Future Work

The current framework establishes the theoretical and computational foundation for integrating physical constraints into LLM-based mechanistic reasoning. Moving forward, we identify several key directions to evolve CMI-AI into a production-ready discovery tool.

5.1 Systematic Empirical Validation

As a primary next step, we intend to conduct a rigorous empirical evaluation of the CMI-AI framework using large-scale reaction datasets such as USPTO [6] and Reaxys [7]. This validation will focus on:

- **Benchmarking Mechanistic Accuracy:** Comparing the success rate of CMI-AI against general-purpose LLMs in predicting correct intermediates for canonical and non-canonical reactions.
- **OOD (Out-of-Distribution) Robustness:** Testing the model on structurally novel substrates to quantify how well the Masking Verification Step (MVS) prevents "stochastic parrot" behavior and improves the calibration of the model's confidence under unfamiliar conditions.

5.2 Expansion of Chemical Scope and Dynamics

While the current model focuses on fundamental thermodynamic parameters (pK_a, BDE), we plan to incorporate more complex chemical phenomena:

- **Stereochemistry and Regioselectivity:** Integrating 3D Graph Neural Networks (GNNs) to process spatial molecular data, enabling the framework to reason about stereochemical outcomes and regioselective preferences.
- **Kinetic Analysis:** Transitioning from "feasibility" to "rate prediction" by incorporating activation energy ($\Delta G^{\ddagger}$) heuristics. This will allow the model to identify the Rate-Determining Step (RDS) more accurately among competing pathways.
- **Catalysis:** Extending the symbolic core to handle transition metal complexes and organometallic catalytic cycles, which require specialized coordination and redox logic.

5.3 Integration with the DMTA Cycle and Self-Driving Labs

Ultimately, CMI-AI is designed to function as an active partner in the Design-Make-Test-Analyze (DMTA) cycle.

- **Closed-Loop Discovery:** We aim to integrate the framework with autonomous synthesis platforms (Self-driving Labs) where AI-suggested mechanisms guide robotic experimentation, and the resulting experimental data (yield, side-products) is fed back into the scoring function for iterative model refinement.
- **Human-AI Collaboration:** Evaluating the impact of CMI-AI in prospective, laboratory-relevant case studies to measure how interpretable mechanistic traces assist human scientists in resolving synthetic bottlenecks.

References

[1] LG AI Research. (2024) EXAONE 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*.

[2] Landrum, G. (2026) RDKit: Open-source cheminformatics. *https://www.rdkit.org*. Accessed: 2026-01-30.

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

[4] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021) On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pp. 610–623.

[5] LangChain Inc. (2026) LangGraph. Software repository. Accessed: 2026-01-30.

[6] Lowe, D. (2017) Chemical reactions from US patents (1976–Sep2016). Figshare dataset. Accessed: 2026-01-30.

[7] Elsevier. (2026) Reaxys. Database. Accessed: 2026-01-30.

[8] Schneider, N., Stiefl, N., & Landrum, G. A. (2016) What's what: The (nearly) definitive guide to reaction role assignment. *Journal of Chemical Information and Modeling* 56(12):2336–2346.

[9] Sutskever, I., Vinyals, O., & Le, Q. V. (2014) Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*.

[10] Weininger, D. (1988) 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.

[11] Wittig, G., & Geissler, G. (1953) Zur Reaktionsweise von Phosphonium-Yliden. *Justus Liebigs Annalen der Chemie* 580:44–57.

[12] Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H., & Laino, T. (2021) Unsupervised attention-guided atom-mapping. *Science Advances* 7(50):eabe4166.

A Appendix / supplemental material

A.1 Dataset: USPTO-50K

We use the USPTO-50K benchmark dataset, a widely adopted, curated subset of chemical reactions extracted from U.S. patents [6]. The dataset originates from the USPTO reaction corpus compiled from patent literature (1976–2016) and was subsequently filtered to a high-quality subset of approximately 50k organic reactions with reliable atom-to-atom mappings and standardized SMILES representations. Following the common "Schneider" formulation [8], USPTO-50K is additionally annotated into 10 coarse-grained reaction classes, which enables controlled benchmarking under both class-conditional and class-agnostic settings.

Representation. Each entry is provided as a reaction SMILES string (reactants and reagents leading to products) with atom mapping indices that specify correspondence between reactant-side and product-side atoms. This mapping is particularly important for our setting because it enables (i) explicit localization of reaction centers (bond changes), (ii) bookkeeping for conservation of atoms through intermediate states, and (iii) structured action parameterization for graph-edit-based mechanistic steps. In our pipeline, we preserve the provided atom mappings and parse all molecules using RDKit, applying standard sanitization and valence checks to ensure graph consistency before use in downstream reasoning and verification components.

Data split. We adopt the standard train/validation/test split used in prior USPTO-50K benchmarks, consisting of 40k reactions for training, 5k for validation, and 5k for testing. This split is commonly used in reaction prediction and retrosynthesis studies, facilitating direct comparison with existing baselines when reporting future empirical results.

Scope and limitations. Although USPTO-50K is attractive due to its curation quality and consistent atom mapping, it represents a relatively narrow slice of chemical space: the benchmark focuses on 10 major reaction families and does not fully capture the long-tail diversity of patent chemistry. As a result, models validated solely on USPTO-50K may overestimate generalization to out-of-distribution mechanisms or under-represent rare condition-dependent pathways. In our future evaluation, we therefore treat USPTO-50K as an initial controlled benchmark and plan to extend validation to broader and more heterogeneous reaction collections.

A.2 Rationale for Selecting EXAONE as the Core Reasoning Model

In this work, we adopt EXAONE [1] as the core neural reasoning engine in CMI-AI because the proposed framework requires a model that can (i) generate chemically structured hypotheses with high instruction fidelity, (ii) maintain long-context coherence across multi-step mechanistic chains, and (iii) support iterative self-correction under externally imposed constraints. Mechanism inference is not a single-shot prediction problem; it is a multi-turn, stateful reasoning process that must preserve consistency between intermediate structures, condition-dependent reactivity assumptions, and the accompanying rationale used for verification. In practice, many general-purpose LLMs exhibit failure modes that are particularly damaging for chemistry—most notably, confident generation of physically implausible steps driven by surface-level correlations, inconsistent bookkeeping of charge/atom mapping across steps, and unstable adherence to required output schemas. These failure modes increase the burden on the symbolic layer and can cause search to waste budget exploring unphysical branches.

EXAONE is well suited for this setting for three operational reasons. First, it shows strong controllability in structured generation: when prompted to output action tuples (step type, reaction-center atoms, bond edits, charge updates, participants, and local justification), the model more reliably conforms to the specified schema, which is essential because CMI-AI executes proposals as explicit graph edits rather than free-form text. High schema compliance directly reduces validator rejections and enables efficient in-loop pruning. Second, EXAONE supports long-horizon coherence required by tree-search-based mechanism inference. As the beam expands, candidates must remain internally consistent with prior intermediate states and condition flags (e.g., protic vs. aprotic solvent, strong base vs. Lewis acid regime). Models that frequently drift in assumptions or "rewrite history" degrade both plausibility scoring and downstream verification; by contrast, a more stable long-context generator improves the quality of partial mechanisms and makes comparisons across branches meaningful.

Third, EXAONE is amenable to verification-driven refinement: the proposed Masking Verification Step (MVS) and retry loop require the model to revise specific causal claims when presented with targeted failure diagnostics (e.g., pK_a directionality mismatch, leaving-group feasibility contradiction, or reaction-center inconsistency). A model that can incorporate structured feedback and produce minimal edits—rather than wholesale rewrites—substantially improves convergence under a limited retry budget.

Importantly, the selection of EXAONE does not imply that chemical validity emerges from the base model alone; rather, EXAONE provides a strong "proposal prior" that is then grounded by deterministic constraints and symbolic checks. Our framework is deliberately modular: EXAONE can be replaced by alternative LLM backbones, but we choose EXAONE because its controllability, coherence, and responsiveness to corrective feedback reduce the overall error rate of candidate generation, thereby allowing the RDKit-based constraint module, the beam-search controller, and MVS to operate more efficiently and transparently. Future comparative studies will evaluate backbone sensitivity by swapping the proposal model while keeping the symbolic and verification components fixed, enabling controlled ablation of how much reliability is attributable to the LLM prior versus the physical grounding and verification layers.

AI Co-Scientist Challenge Korea Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [N/A] .
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[N/A] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "AI Co-Scientist Challenge Korea paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that our contribution is a physically grounded hybrid inference framework (LLM+RDKit), beam-search-based mechanism exploration, and masking-based verification (MVS), which matches the scope of the paper (Method in Sec. 4).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We explicitly note that empirical validation and broader benchmarking are deferred to future work, and we discuss dataset scope limitations (e.g., USPTO-50K coverage) and expected out-of-distribution risks (Dataset/Scope paragraph and Future Work statement).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper provides algorithmic formulations and scoring objectives, but does not present formal theorems or proofs requiring a complete assumption/proof treatment.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: This submission focuses on method design and verification protocol; we do not report completed experimental results that require reproduction at this stage.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: We do not release implementation code or experimental artifacts in this submission because we do not report finalized experiments; we describe the dataset and pipeline components to support future reproducible evaluation.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [N/A]

   Justification: Since we do not report finalized experiments, there are no training hyperparameters, optimizer settings, or evaluation protocols to disclose beyond the dataset description.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: No experimental results (thus no error bars or significance tests) are reported in the current version.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [N/A]

Justification: No experimental runs are reported, so compute resources and runtime statistics are not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://nips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work uses publicly available reaction benchmarks and standard open-source chemistry tooling, does not involve human subjects, and is conducted in accordance with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The current manuscript does not yet include a dedicated broader impacts discussion; we will add a short section covering potential benefits (accelerating mechanism analysis) and risks (dual-use chemical knowledge, incorrect mechanistic guidance) along with mitigation considerations.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release a new pretrained model, weights, or a scraped dataset as part of this submission, so release-time safeguards are not applicable in the current form; future releases (if any) will consider controlled access and usage guidelines.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: While we cite and use USPTO-50K and RDKit, the manuscript does not yet explicitly enumerate licenses/terms of use for each external asset; we will add an explicit license/terms statement for the dataset and software dependencies in the final version.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

Justification: The paper does not introduce or release new datasets, model checkpoints, or code artifacts as part of this submission.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve crowdsourcing or research with human participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human-subject study is conducted, so IRB approval is not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.