

INDIBATOR: Diverse and Fact-Grounded Individuality for Multi-Agent Debate in Molecular Discovery

Anonymous Authors¹

Abstract

Multi-agent systems have emerged as a powerful paradigm for automating scientific discovery. To differentiate agent behavior in the multi-agent system, current frameworks typically assign generic role-based personas such as “reviewer” or “writer” or rely on coarse grained keyword-based personas. While functional, this approach oversimplifies how human scientists operate, whose contributions are shaped by their unique research trajectories. In response, we propose INDIBATOR, a framework for molecular discovery that grounds agents in individualized scientist profiles constructed from two modalities: publication history for literature-derived knowledge and molecular history for structural priors. These agents engage in multi-turn debate through proposal, critique, and voting phases. Our evaluation demonstrates that these fine-grained individuality-grounded agents consistently outperform systems relying on coarse-grained personas, achieving competitive or state-of-the-art performance. These results validate that capturing the “scientific DNA” of individual agents is essential for high-quality discovery.

1. Introduction

Large language models (LLMs) have demonstrated remarkable performance across a wide variety of tasks (Singh et al., 2025; Anthropic, 2024; Team et al., 2025; DeepSeek-AI et al., 2025). Beyond direct prompting, recent works have introduced AI agents capable of planning and executing actions over multiple iterations (Yao et al., 2023; Schick et al., 2023; M. Bran et al., 2024). While impressive, single-agent systems often encounter constraints such as bounded context windows and limited perspective diversity. To address

this, multi-agent systems have emerged as a powerful paradigm for automated discovery (Du et al., 2024; Liang et al., 2024; Chan et al., 2024). By leveraging collaborative intelligence, these systems effectively simulate the real-world research process with growing applications in scientific discovery (Lu et al., 2024; Du et al., 2025; Gottweis et al., 2025) and molecular discovery (Kim et al., 2025).

To differentiate the conversational behavior of each agent, prior works typically assign distinct personas through role-play prompting (Kong et al., 2024; Zhou et al., 2024; Park et al., 2023; Piao et al., 2025), e.g., “planner”, “verifier”, or “reviewer”, or through keywords (Su et al., 2025). While this role-based or keyword-based separation effectively shapes output style, it often oversimplifies the rich reality of how human scientists operate. In practice, a scientist’s individuality is defined not merely by a coarse-grained generic role or a set of keywords, but by their unique fine-grained research trajectory, a distinctive “scientific DNA” composed of cumulative experiences and domain-specific intuitions. By ignoring this, current systems fail to leverage the deep, nuanced insights characteristic of real-world collaboration.

The existence of such scientific DNA is particularly well established in the domain of drug discovery. Chemists exhibit distinctive styles for designing new molecules, such as preferences for particular scaffolds, functional groups, and reaction motifs (Pedreira et al., 2019; Choung et al., 2023), based on their own research trajectory. Recently, Blevins & Quigley (2025) quantified this phenomenon, demonstrating that models can identify which of 1,815 chemists synthesized a molecule with 60% top-5 accuracy from structure alone. While they frame this as “Clever Hans” (Lapuschkin et al., 2019) leakage problem that distorts benchmark evaluations, we reinterpret it as a *blueprint for agent design*. We argue that these styles encode heuristics for effectively navigating chemical space and representing the expertise diversity that mimics real-world collaboration.

In response, we propose INDIBATOR, a multi-agent framework for molecular discovery that bridges the gap between generic coarse-grained personas and chemical reality by grounding agents in individual research trajectories, as illustrated in Figure 1. Instead of relying on heuristically predefined roles or keywords, INDIBATOR constructs agent

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

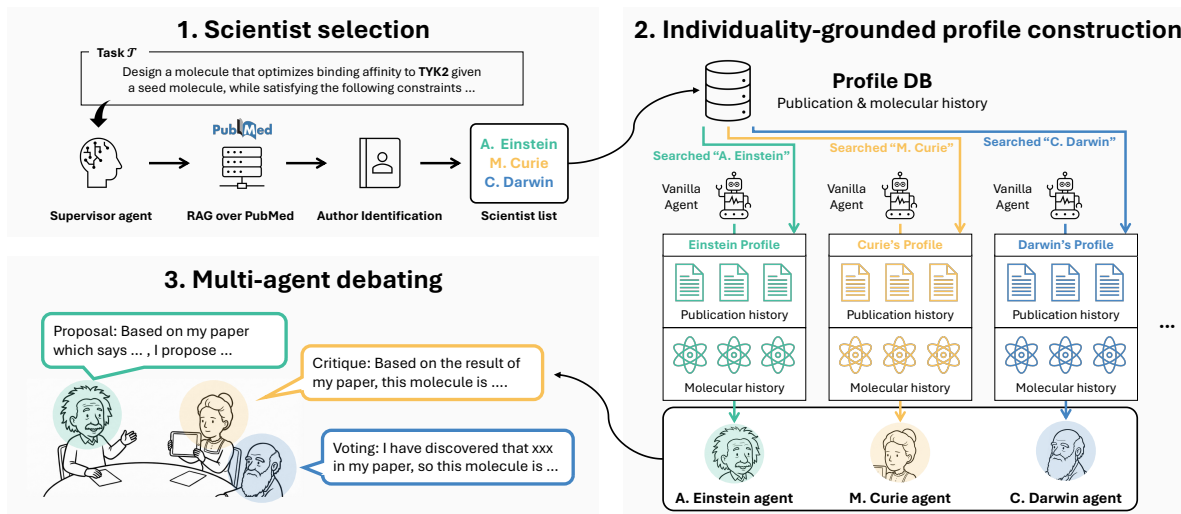


Figure 1. Overview of INDIBATOR. Given a task, the supervisor agent selects relevant scientists by identifying the authors of publications by RAG. Next, individuality is grounded for each agent with scientist profiles, consisting of publication history and molecular history of each scientist. Finally, multi-agents debate to iteratively generate candidate molecules with proposal, critique, and voting phases.

profiles utilizing two informative sources that encode research trajectory: (1) *publication history*, a collection of publications that define the agent’s literature-derived knowledge and methodological preferences, and (2) *molecular history*, a set of previously developed molecules that establishes structural priors, such as preferred scaffolds and functional groups.

This data-driven profile provides unique individuality to each agent, effectively mirroring the real-world scientific process where discovery emerges from researchers’ unique cumulative knowledge and inductive biases. This individuality-driven design provides two key benefits: (1) *diversity*, where unique agent profiles prevent redundant reasoning among the agents; and (2) *fact-grounding*, where explicit reliance on publication and molecular records empowers reasoning grounded in verifiable evidence.

To demonstrate the effect of our fine-grained individuality, we implement a multi-agent debating system consisting of three iterative phases: (1) proposal, (2) critique, and (3) voting. During these phases, each agent proposes molecular candidates, critiques proposals, and assigns scores based on their specific expertise, mirroring the collective intelligence of real-world scientist teams.

We empirically evaluate INDIBATOR across three downstream tasks: protein-conditioned molecule generation, bioactivity-guided molecule generation, and goal-directed lead optimization. Our results show that INDIBATOR consistently outperforms vanilla debating systems and achieves competitive or state-of-the-art performance across benchmarks. Moreover, we provide comprehensive analyses demonstrating the impact of individuality, which validates that capturing the nuanced scientific DNA is a fundamental driver of molecular design. While the principle of expertise-

grounded individuality may generalize to other scientific domains, our results demonstrate that it is a critical component for enhancing molecular discovery, where scientist style provides concrete empirical grounding.

2. INDIBATOR

The INDIBATOR framework instantiates a collective of scientist agents grounded in their own unique research trajectories. Unlike conventional multi-agent systems (Kong et al., 2024; Zhou et al., 2024; Park et al., 2023), our approach ensures more fine-grained individuality of each agent with research trajectory profiles. Specifically, we condition each agent on a distinct real-world profile derived from their prior publications and historical molecular discoveries. This provides two key benefits:

- **Diversity:** Each agent’s system prompt is uniquely constructed based on their expertise, preventing redundant reasoning across the multi-agent ensemble.
- **Fact-grounding:** Each agent’s reasoning is grounded in real-world profiles, ensuring that their arguments are supported by concrete empirical evidence, including papers and discovered molecules.

2.1. Individuality-grounded Profile Construction

To ensure the individuality of each scientist agent, we construct expertise profiles from two modalities: (1) *publication history*: a collection of publications retrieved from PubMed (Luna, 2024; Cho, 2024) that define the agent’s literature-derived knowledge, including their research focus and methodological preferences; and (2) *molecular history*: a set of molecules previously developed by the scientist

that establishes structural priors, represented as SMILES strings (Weininger, 1988). Notably, the inclusion of molecular history M_i is motivated by the Clever Hans phenomenon in chemistry (Blevins & Quigley, 2025), which highlights a correlation between molecular structures and the scientists associated with their discovery.

In detail, given a research objective or task description \mathcal{T} , a supervisor agent employs retrieval-augmented generation (RAG; Lewis et al., 2020) over a vector space of literature in PubMed to identify the most relevant research papers. The supervisor then extracts the first and last authors to represent the primary researchers and principal investigators, respectively. These identified scientists construct a set of scientist agents, $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$, where N is a hyperparameter that defines the number of scientists.

Finally, each scientist agent s_i is initialized with an expertise profile $E_i = \{P_i, M_i\}$, where P_i denotes their publication history, including the titles and abstracts, and M_i denotes the molecular history, i.e., a set of molecules discovered by the scientist. Specifically, a subset of publications is selected based on the frequency of task-relevant keywords, while molecules are retrieved based on their structural similarity to a provided seed molecule if available.

2.2. Multi-agent Debating System

The proposed debating system consists of three phases: proposal, critique, and voting. This iterative process continues until reaching a maximum round limit or accumulating a sufficient number of candidates. We provide a detailed qualitative case study of a diversified and fact-grounded debating process in Figure 3 and detailed prompts in Appendix A.

Proposal. During the proposal phase, each *scientist agent*, i.e., an agent conditioned with individual scientist profiles, generates molecular candidates grounded in their specific expertise. Each agent is provided with a prompt including their expertise profile E_i and the task description \mathcal{T} . Scientists propose k candidates, accompanied by rationales that link each proposal to their prior knowledge. In subsequent rounds, agents also receive candidates and critiques from previous iterations to facilitate refinement.

Critique. The critique phase operates in two stages to generate feedback for candidate molecules. First, in an optional self-critique stage, scientist agents invoke tools to evaluate their own proposals, identifying weaknesses and modifying their designs. Next, agents engage in a cross-critique stage, where they evaluate peer proposals to simulate a collaborative review process similar to the real-world scientific discovery. In this step, agents leverage their own personas to suggest domain-specific modifications, thereby ensuring that final candidates are robust across multiple criteria.

Voting. In the voting phase, scientists assess the candidates, incorporating the insights from the critique phase. Each scientist agent s_i evaluates the candidate pool, assigning a scalar score $s \in [0, 1]$ based on three objectives: task relevance, synthetic feasibility, and novelty. Based on these scores, each agent casts votes for the top t candidates. These votes are subsequently aggregated to determine the global ranking. The highest-ranked candidates either proceed to the subsequent round or are selected as the final candidates.

3. Downstream Task Evaluation

Here, we evaluate the effectiveness of our proposed INDIBATOR, on three molecular downstream tasks: (1) protein-conditioned molecule generation (Section 3.1), (2) bioactivity-guided molecule generation (Section 3.2), and (3) goal-directed lead optimization (Section 3.3). Notably, we utilize the Deepseek-V3.2 (DeepSeek-AI et al., 2025) backbone. We provide task prompts in Appendix A and further experimental settings in Appendix B.

3.1. Protein-conditioned Molecule Generation

Task description. The goal of the protein-conditioned molecule generation task is to generate molecules with a high binding affinity to the target protein. In detail, we select eight target proteins: TYK2, JNK1, CDK2, P38, CA2, THROMBIN, FABP4, and DHFR. The first four proteins are selected following the binding affinity prediction task of Boltz-2 (Passaro et al., 2025). Since these targets are all kinases, to expand the scope of evaluation, we expanded the evaluation scope to include non-kinase targets: CA2 (metalloenzyme), THROMBIN (serine protease), FABP4 (lipid-binding protein), and DHFR (oxidoreductase). This ensures a diverse coverage of protein domains and ligand interaction mechanisms. We generate 1,000 candidate molecules for each protein.

We employ two types of metrics: binding affinity and diversity. First, for the binding affinity, we utilize Boltz-2 (Passaro et al., 2025) as our proxy. Specifically, we employ `affinity_pred_value`, which quantifies the specific affinity of various binders and tracks how these values change in response to small molecular modifications. For standardized comparison, we convert the binding affinity value, represented as $\log_{10}(\text{IC}_{50})$, into kcal/mol. Our final evaluation metrics consist of (1) the Top-1 binding affinity and (2) the mean of the Top-10 binding affinities achieved by the 1,000 generated candidates.

In addition, we provide two metrics to validate whether individuality matters for the diversity of generations. In detail, the diversity is computed with (3) internal diversity (Int-Div; Polykovskiy et al., 2020) and (4) the number of cir-

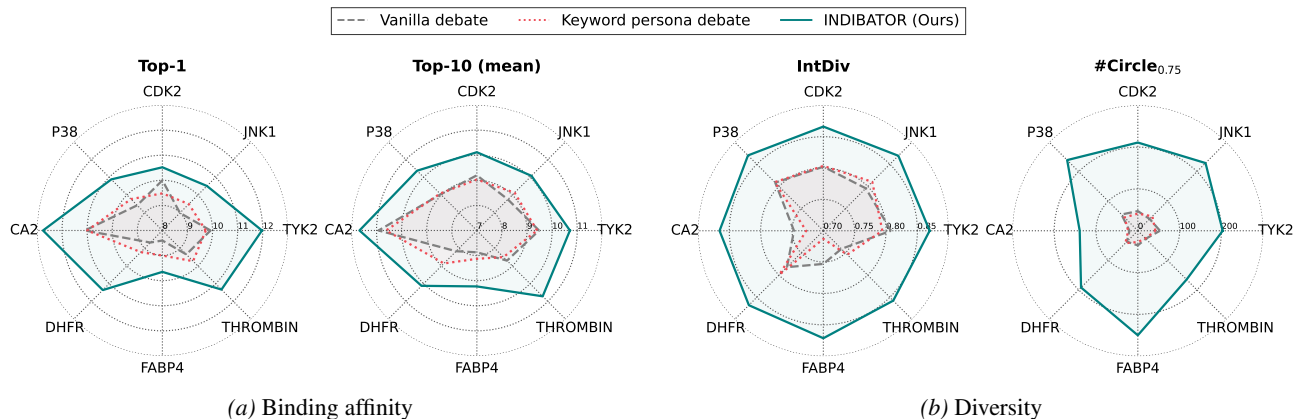


Figure 2. Results of protein target molecular generation. The left and right panels illustrate the docking scores and diversity of molecules, respectively. The gray, red, teal colors denote vanilla debate, keyword persona debate, and INDIBATOR (ours), respectively. Notably, the docking scores are presented in absolute values, with higher scores representing superior binding.

cles (#Circles; Xie et al., 2023) following Jang et al. (2024). IntDiv measures the average pairwise Tanimoto similarity of molecules while #Circles_h computes the number of mutually exclusive circles where each circle is constructed with the Tanimoto similarity threshold $h = 0.75$.

Baselines. We establish VANILLADEBATE and KEYWORDDEBATE as our baselines. Both baselines follow the identical debating framework of our INDIBATOR. However, VANILLADEBATE operates without any profile, while KEYWORDDEBATE constructs the profile with research keywords extracted from the publication history of each agent, inspired by VirSci (Su et al., 2025).

Results. We provide the results in Figure 2 and detailed values in Appendix C. INDIBATOR consistently beats the baselines across all target proteins and metrics, demonstrating superior performance in both binding affinity and molecular diversity. Importantly, while the baselines frequently suffer from mode collapse, indicated by their lower diversity, INDIBATOR successfully navigates the chemical space to generate a high volume of structurally distinct clusters without compromising optimization quality.

Notably, the KEYWORDDEBATE demonstrates negligible improvement compared to the VANILLA baseline, highlighting that coarse-grained keywords are insufficient for the nuanced reasoning needed for molecular design. Our superior performance across diverse protein families validates that our fine-grained profile-grounded agents can effectively adapt their search strategy to distinct biological interactions.

3.2. Bioactivity-guided Molecule Generation

Task description. This task aims at maximizing molecular biological activity properties under unconstrained conditions. It includes three bioactivity optimization tasks:

Table 1. Results of PMO-1K benchmark. We mark the best results in **bold**. Method denoted with an asterisk (*) indicated LLM-based baselines implemented by the authors to ensure consistent experimental settings.

	Model	GSK3 β	DRD2	JNK3
Structure	GP BO	0.611	0.857	0.346
	REINVENT	0.589	0.775	0.315
	Genetic GFN	0.637	0.809	0.409
	Graph GA	0.523	0.833	0.301
	Aug. Mem.	0.539	0.795	0.294
LLM	LICO-L	0.617	0.859	0.336
	MOLLEO-B	0.397	0.910	0.186
	MOLLEO-D	0.496	0.812	0.342
	VANILLA*	0.419	0.921	0.310
	VANILLADEBATE*	0.477	0.902	0.161
	KEYWORDDEBATE*	0.449	0.929	0.185
	MT-Mol	0.308	0.756	0.125
	INDIBATOR (Ours)	0.942	0.950	0.914

GSK3 β , DRD2, and JNK3. In detail, these properties are:

- GSK3 β : Inhibition of glycogen synthase kinase-3 β .
- DRD2: Binding affinity for dopamine type 2 receptor.
- JNK3: Inhibition of c-Jun N-terminal kinase-3.

Following prior works (Kim et al., 2025; Nguyen & Grover, 2025), we conduct experiments on the practical molecular optimization (PMO)-1K benchmark (Gao et al., 2022), which computes the score among 1,000 generated molecules. Notably, we excluded other tasks in PMO, such as rediscovery of a given molecule or isomer generation that satisfies the molecular formula. These tasks represent arithmetic structural puzzles where success depends on precise reconstruction instead of broad exploration of diverse molecules. For the metric, we report average of top-10 AUC scores, the area under the curve (AUC) of the top-10 average performance versus oracle calls.

Table 2. Goal-directed lead optimization results. Teal highlights the improvement to the VANILLADEBATE. Larger absolute values denote better binding.

	Target protein	parp1			fa7			5ht1b			braf			jak2		
	Seed score	-7.3	-7.8	-8.2	-6.4	-6.7	-8.5	-4.5	-7.6	-9.8	-9.3	-9.4	-9.8	-7.7	-8.0	-8.6
Learning-based	Graph GA	-8.3	-8.9	-	-7.8	-8.2	-	-11.7	-12.1	-	-9.8	-	-11.6	-8.7	-9.2	-
	RetMol	-9.0	-10.7	-10.9	-8.0	-	-	-12.1	-9.0	-	-	-11.6	-	-8.2	-9.0	-
	GenMol	-10.6	-11.0	-11.3	-8.4	-8.4	-	-12.9	-12.3	-11.6	-10.8	-10.8	-10.6	-10.2	-10.0	-9.8
Inference-only	VANILLA*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	VANILLADEBATE*	-12.8	-11.6	-9.7	-7.8	-7.0	-	-12.3	-10.6	-10.1	-10.2	-9.6	-10.2	-9.8	-9.6	-8.5
	INDIBATOR (Ours)	-12.1	-11.5	-16.7	-9.2	-7.0	-	-12.4	-11.6	-10.5	-10.6	-9.8	-10.5	-9.8	-11.0	-8.8

Baselines. We benchmark against eleven baselines, categorized into five structure-based and six LLM-based approaches. The structure-based approaches includes GP BO (Srinivas et al., 2010), REINVENT (Olivecrona et al., 2017), Genetic GFN (Kim et al., 2024), Graph GA (Jensen, 2019), and Augmented Memory (Guo & Schwaller, 2024). Additionally, we compare against LLM-based approaches, including LICO (Nguyen & Grover, 2025), MOLLEO (Wang et al., 2025), the role-based multi-agent system MT-Mol (Kim et al., 2025), VANILLA, VANILLADEBATE, and KEYWORDDEBATE inspired by VirSci (Su et al., 2025). Here, VANILLA indicates the LLM prompting approach without any debate. Notably, MT-Mol is a critical baseline for evaluating the performance of a multi-agent system with generic role-based assignment.

Results. We report the results in Table 1. We observe that INDIBATOR consistently enhances the performance across all tasks. It is notable that our method demonstrates a substantial performance margin over the role-based MT-Mol and KEYWORDDEBATE. This validates that fine-grained, diverse, and fact-grounded individuality provides a more effective inductive bias for chemical space navigation than generic role prompts or keywords. Furthermore, INDIBATOR outperforms optimization baselines such as Genetic GFN by significant margins, ranging from 17.4% (DRD2) to 123.5% (JNK3) across the evaluated targets.

3.3. Goal-directed Lead Optimization

Task description. The goal of the goal-directed lead optimization task is to generate leads given an initial seed molecule. The leads are the molecules that exhibit improved target properties while maintaining the similarity with the given seed molecule. Following Lee et al. (2025), the objective is to maximize the binding affinity measured by the docking score while satisfying the following constraints: $QED \geq 0.6$, $SA \leq 4$, and $sim \geq 0.6$. The similarity is defined as the Tanimoto similarity between the Morgan fingerprints of the generated and seed molecules. We adopt five target proteins: parp1, fa7, 5ht1b, braf, and jak2, and each protein includes three different seed molecules. We evaluate performance based on the docking score of the most optimized lead.

Baselines. We compare against five baselines categorized into two paradigms. The first category comprises learning-based optimization methods: Graph GA (Jensen, 2019), RetMol (Wang et al., 2023), and GenMol (Lee et al., 2025). Crucially, these models are explicitly trained with feedback loops to satisfy constraints and maximize docking scores, establishing a strong performance standard. In contrast, INDIBATOR operates without any task-specific fine-tuning. Following this, the second category consists of inference-only LLM baselines, VANILLA and VANILLADEBATE, which follow the settings in the previous experiments.

Results. We provide the results in Table 2. While the VANILLA baseline fails to generate a single qualified molecule that satisfies all the constraints, debate showed improvement. However, while VANILLADEBATE shows improvement, INDIBATOR consistently generates more optimized leads, demonstrating that our expertise-grounded profiles provide the appropriate guidance to navigate constrained chemical space. While INDIBATOR does not uniformly surpass state-of-the-art baselines and shows only competitive results, this is expected as the baselines are trained to maximize target properties while our method operates solely on inference-time and does not include any task-specific training required by the baselines.

4. Analysis

In this section, we conduct a comprehensive analysis to dissect the mechanisms behind INDIBATOR’s performance utilizing the bioactivity-guided molecule generation task. We begin by presenting qualitative results in Section 4.1, illustrating how agents leverage their unique research trajectories for reasoning. Next, we investigate the impact of individuality by addressing three key research questions:

- **Granularity** (Section 4.2): Does the granularity of the profile impact the performance?
- **Diversity** (Section 4.3): Does the performance stem from the heterogeneity of expert perspectives?
- **Fact-Grounding** (Section 4.4): Is grounding agents in real-world data essential compared to synthetic or hallucinated profiles?

Task \mathcal{T}

Your task is to design a SMILES string for a molecule that satisfies the following condition:

- Design a drug-like molecule with high predicted JNK3 (c-Jun N-terminal kinase 3) inhibitory activity.
- Consider structural features known to enhance kinase inhibition.

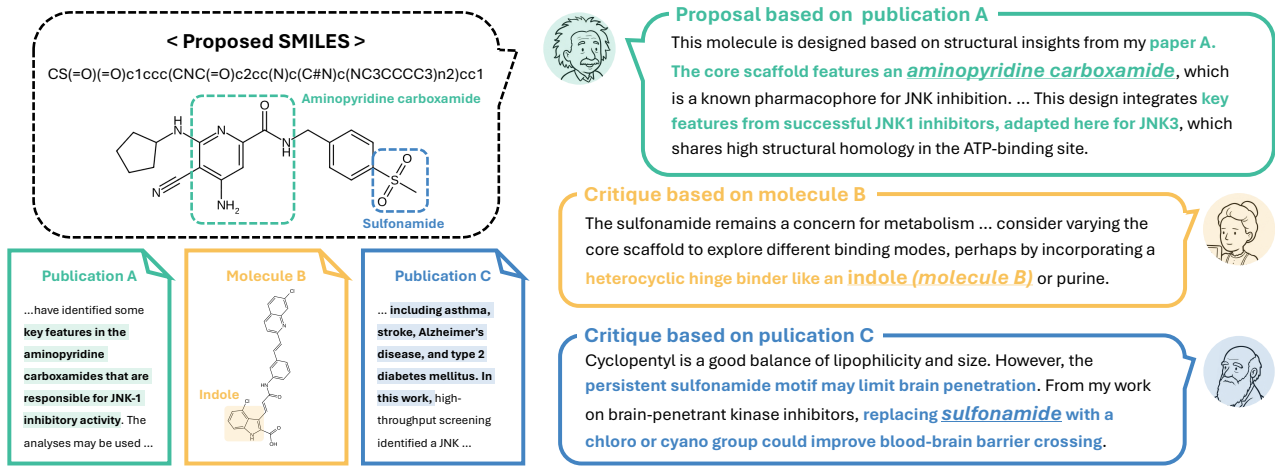


Figure 3. Qualitative case study on individuality grounded agents. We provide a qualitative analysis of the JNK3 inhibition guided molecule generation task. Specifically, we show how an agent leverages prior publications and molecules to propose a candidate, while other agents utilize their profiles to offer targeted critiques.

Finally, we provide an ablation study in Section 4.5 to evaluate the impact of the number of scientists and each component in INDIBATOR.

4.1. Qualitative Case Study

We illustrate how grounding agents in distinct publication and molecular histories shapes their reasoning during the JNK3 inhibition guided molecule generation task in Figure 3. This shows that grounding agents in their individual scientific profiles leads to distinct, chemically plausible reasoning trajectories. It is notable that we present partial examples for simplicity, while all agents engage in the debate in parallel for the entire proposal, critique, and voting phases. We provide more detailed examples in Appendix C.

The **first agent**, retrieving its prior publication on 3D quantitative structure–activity relationship (QSAR) on JNK1 inhibitors (Yi & Qiu, 2008), proposes a candidate molecule with an *aminopyridine carboxamide* scaffold. This is scientifically sound considering the high structural homology between the JNK1 and JNK3 ATP-binding pockets, transferring the pharmacophore is a logical exploitation (Liu et al., 2006). This proposal triggers a structural critique from a **second agent**, which proposes to refine the scaffold driven by its background on indolin-2-one Aurora B inhibitors (Zhang et al., 2015). To be specific, this critique is to replace the core scaffold, which explores novel binding modes with an indole core (Chen et al., 2016). Finally, the **third agent** critiques to refine the molecule for the specific therapeutic indication. Grounded in a central nervous system (CNS) focused publication (Zheng et al., 2014), the agent correctly

identifies the sulfonamide motif as a blood-brain barrier (BBB) liability due to its high polarity. It suggests replacing the sulfonamide with the chloro or cyano groups, which reduces the polar surface area (Kelder et al., 1999).

4.2. Effect of Granularity

Here, we analyze the effect of granularity, which refers to the depth and specificity of information used to construct an agent’s persona, ranging from generic role assignments to detailed research trajectory-based profiles.

To analyze this, we consider three baselines representing the spectrum of granularity: VANILLADEBATE, role persona, and keyword persona. In detail, role persona represents coarse-grained individuality, where agents are assigned based on generic, LLM-generated task-related roles (e.g., medicinal chemist, cheminformatics scientist, etc.). Next, the keyword persona represents mid-level granularity where agents are defined by keywords extracted from the publication history of each agent, inspired by VirSci (Su et al., 2025). Finally, INDIBATOR represents our fine-grained individuality based on publication and molecular histories.

The results in Table 3 demonstrate that increasing profile granularity consistently improves the performance. While role persona and keyword persona offer marginal gains over the VANILLADEBATE, INDIBATOR, which utilizes the full publication and molecular history, significantly outperforms all baselines in terms of both performance and diversity. This confirms that capturing the nuanced “scientific DNA” rather than just generic roles or keywords is critical for

Table 3. **Comprehensive quantitative analysis.** We evaluate models across three perspectives: **Granularity** (level of detail in profile), **Diversity** (heterogeneity of agents), and **Fact-grounding** (relevance and truthfulness of knowledge). $-$, \blacktriangle , and \bullet indicates the low-, mid-, and high-level of each property, respectively. Higher values are better across all metrics and best results are highlighted in **bold**.

Model	Agent Properties			GSK3 β				DRD2				JNK3			
	Gran.	Div.	Fact.	AUC	IDiv	#C. ₇₅	#C. ₈₅	AUC	IDiv	#C. ₇₅	#C. ₈₅	AUC	IDiv	#C. ₇₅	#C. ₈₅
<i>Baseline</i>															
VANILLADEBATE	$-$	$-$	$-$	0.477	0.816	48	8	0.902	0.835	57	10	0.161	0.809	55	9
<i>Ablation on Granularity</i>															
Role persona	$-$	\bullet	$-$	0.625	0.816	54	9	0.933	0.823	56	11	0.178	0.812	60	6
Keyword persona	\blacktriangle	\bullet	\bullet	0.449	0.813	47	10	0.929	0.832	52	10	0.185	0.808	52	6
<i>Ablation on Diversity</i>															
Single-profile	\bullet	$-$	\bullet	0.285	0.734	20	5	0.857	0.809	31	9	0.147	0.787	39	8
Massive single-profile	\bullet	$-$	\bullet	0.559	0.829	96	21	0.950	0.750	19	6	0.453	0.767	34	8
<i>Ablation on Fact-Grounding</i>															
LLM-generated profile	\bullet	\bullet	$-$	0.501	0.792	44	7	0.927	0.813	52	9	0.235	0.799	48	7
Random-profile	\bullet	\bullet	\blacktriangle	0.884	0.850	125	23	0.929	0.833	78	22	0.334	0.837	117	25
INDIBATOR (Ours)	\bullet	\bullet	\bullet	0.942	0.850	182	35	0.950	0.833	78	21	0.914	0.843	115	26

navigating chemical spaces effectively.

4.3. Effect of Diverse Agents

Here, we analyze the effect of diversity to observe whether the collaborative performance stems merely from the aggregation of knowledge or from the interaction between diverse and heterogeneous expert perspectives.

We consider three baselines: VANILLADEBATE, single-profile, and massive single-profile. By assigning identical (or null) profiles across all agents in baselines, we can explicitly observe the impact of diversity. In detail, single-profile forces multiple agents to share an identical profile (non-diverse profile), where the profile is selected as the most relevant for the scientists. Next, in a massive single-profile, every agent is assigned an identical and comprehensive profile constructed from the union of scientist profiles in INDIBATOR. Due to context window constraints, this union aggregates 50% of the selected profiles.

The results in Table 3 highlight the necessity of collaboration between diverse agents. Single-profile performs even worse than the VANILLADEBATE in most metrics, suggesting that enforcing a narrow and homogeneous perspective hinders exploration. Crucially, INDIBATOR outperforms the massive single-profile, proving that performance gains stem not merely from the volume of knowledge, but from the diversity of perspectives that enables agents to cross-examine and propose diverse candidates.

4.4. Effect of Fact-grounding Agents

To analyze the importance of grounding agents in factual data, we consider three baselines: VANILLADEBATE, LLM-generated profile, and random-profile. In this experiment, we ensure that every agent possesses a unique profile to decouple the benefits of fact-grounding from those of diversity. In detail, the agents in LLM-generated profiles are

initialized with synthetic publication and molecular histories generated by LLMs given a scientist name. Although these profiles possess the same structure of real profiles, they cannot guarantee the factual accuracy due to the potential hallucinations. The random-profile assigns agents complete and factual but task-irrelevant profiles.

The results in Table 3 indicate that fact-grounding is significant. Specifically, the LLM-generated profile performs poorly, often inferior to the random-profile baseline. This suggests that hallucinated expertise introduces noise that degrades reasoning more than irrelevant but real expertise. INDIBATOR, which grounds agents in actual publication and molecular history, achieves superior performance across all benchmarks. This shows that our profiles provide true inductive biases grounded in established knowledge, effectively guiding exploration toward biologically plausible regions.

4.5. Ablation Study

The number of scientists. To analyze the impact of the number of scientists who engage in the debate, we evaluate the performance of INDIBATOR and VANILLADEBATE while varying the number of scientist agents (N) from 5 to 50 on the JNK3 inhibition guided molecule generation task. We provide the results in Figure 4.

The results demonstrate that increasing the number of scientists (N) enhances the performance in INDIBATOR while the VANILLADEBATE exhibits performance degradation due to the reduced number of rounds. Specifically, as the number of scientists increases, the debate concludes in a single round for both models. For INDIBATOR, this is sufficient as the diverse agents generate high-quality initial proposals that cover the chemical space effectively. In contrast, the VANILLADEBATE relies on iterative debate process. This highlights that individuality enables a more efficient scaling law, where expanding the diversity of perspectives can

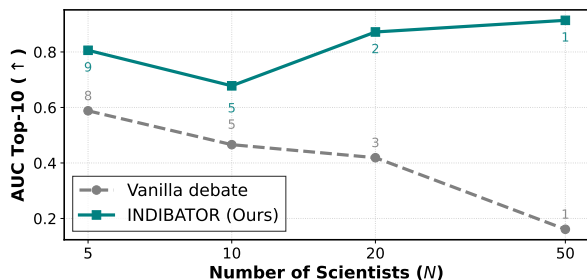


Figure 4. **Effect of the number of collaborators.** Annotated numbers above and below each data point indicate the number of debate rounds required to generate 1,000 candidates.

effectively substitute for the iterative debate.

Effect of each component. To evaluate the individual contributions of each phase within our framework, we conduct an ablation study by removing the critique and voting phases, and individuality (VANILLADEBATE). We evaluate the performance on all three bioactivities. As a result, as shown in Table 4, we observe a performance improvement as each component is integrated. Notably, the most critical increment occurs with the individuality, identifying it as the dominant factor contributing to INDIBATOR’s success.

5. Related Work

LLM-based multi-agent systems. The AI agents have evolved rapidly from single-agent frameworks (Yao et al., 2023; Schick et al., 2023; Shinn et al., 2023) to multi-agent systems (MAS) that leverage collaborative intelligence (Du et al., 2024; Liang et al., 2024; Chan et al., 2024; Lu et al., 2024; Mitchener et al., 2025). By assigning distinct personas to LLMs, MAS can simulate complex interactions, effectively leveraging each agent’s capabilities and expertise. Standard approaches typically employ role-play prompting (Kong et al., 2024; Zhou et al., 2024; Park et al., 2023; Piao et al., 2025) to instantiate generic personas by prompting (e.g., “You are an expert in biology.”). However, this coarse-grained role-based agent scales poorly to massive multi-agent scenarios, where defining a sufficient number of distinct and specialized roles becomes intractable.

Recent works have attempted to mitigate this by introducing more fine-grained personas, such as keywords. For instance, VirSci (Su et al., 2025) constructs scientist agents using keywords based on publications and collaboration networks, demonstrating effective multi-agent collaboration for paper abstract generation. However, keywords alone lack the granular expertise required for real-world scientific debate. Moreover, in the molecular discovery domain, literature-derived knowledge is insufficient as chemists exhibit distinctive structural priors such as scaffolds and functional groups, which are not fully captured by publication texts alone. To address this, INDIBATOR incorporates both pub-

Table 4. Ablation study on each component

Setting	GSK3 β	DRD2	JNK3
(-) individuality	0.477	0.902	0.161
(-) critique&vote	0.863	0.946	0.630
INDIBATOR	0.942	0.950	0.914

lication and molecular histories into each agent’s profile. This enhances individuality, promoting both diverse and fact-grounded collaboration among agents.

LLM in molecular discovery. Recent advancements have increasingly adapted LLMs for molecular discovery. Approaches like LICO (Nguyen & Grover, 2025) and MOLLEO (Wang et al., 2025) extend LLMs with structured embeddings or evolutionary search to enable molecule generation. To address the complex reasoning in the molecular domain, a few works have evolved into agent systems. For instance, ChemCrow (M. Bran et al., 2024), a single-agent system that combines general-purpose LLMs with chemistry tools and a ReAct-based reasoning loop, and MT-Mol (Kim et al., 2025), a multi-agent system that operates tool-guided reasoning and role-specialized LLM agents.

However, a critical limitation of these frameworks comes from their reliance on single-agent architectures or coarse-grained role-based personas, rather than grounded individual expertise. Existing agents lack the rich context of a scientist’s research trajectory, such as the prior publications and previously discovered molecules that define their unique inductive bias. INDIBATOR addresses this by explicitly grounding each agent in a comprehensive profile of their actual research trajectory, effectively replacing generic role-play with collaboration driven by distinct scientific DNA.

6. Conclusion

We presented INDIBATOR, a multi-agent framework that alters coarse-grained generic role-playing by grounding scientist agents in their unique research trajectories. By constructing individual profiles from publication and molecular history for each agent, the system initiates agents with a distinct “scientific DNA” that guides their knowledge-grounded reasoning. Our evaluation across diverse molecular discovery tasks demonstrates that this individuality-based approach consistently outperforms vanilla debating systems and achieves competitive or state-of-the-art performance compared to other baselines. Furthermore, we empirically validated the three-fold benefits of our framework, i.e., granularity, diversity, and fact-grounding, confirming that capturing the nuanced inductive biases of individual researchers is a critical component for high-quality scientific discovery. We believe that our framework establishes a foundation for incorporating broader modalities, such as conversation records, further enhancing the fidelity of agents in domain-specific environments.

Impact Statement

This paper presents a work of using multi agents for molecular discovery. By simulating realistic scientific debate through agents grounded by individual research trajectories, this framework aims to significantly accelerate the drug design pipeline and improve the factual reliability of AI-driven scientific discovery. While our framework currently optimizes drug-likeness and synthetic accessibility, future open-source releases or deployments should involve safety guardrails to prevent the targeted design of harmful compounds.

Ethical Consideration

We acknowledge the ethical considerations regarding scientist profiles in this work. This work uses publicly available academic records, including titles, abstracts, and molecular discoveries from PubMed, to construct expertise profiles for large language model (LLM) agents. The proposal, critiques, and voting generated by these agents are outcomes of the LLM’s probabilistic generation and do not represent the actual opinions, unpublished insights, or endorsement of the real-world scientists cited. Resemblance to the actual private reasoning of individuals is a result of the model’s grounding in their public work. The use of specific scientist profiles in this study is strictly for the purpose of validating the efficacy of individuality grounding in molecular discovery. Additionally, our system could be misused for unethical purposes, such as automating the creation of toxic or harmful molecules. To mitigate these risks, future work should explore safeguards and establish ethical guidelines.

References

- Anthropic. Introducing claude 4, 2024. URL <https://www.anthropic.com/news/claude-4>. 1
- Blevins, A. D. and Quigley, I. K. Clever hans in chemistry: Chemist style signals confound activity prediction on public benchmarks, 2025. URL <https://arxiv.org/abs/2512.20924>. 1, 3
- Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=FQepisCUWu>. 1, 8
- Chen, H., Yang, H., Wang, Z., Xie, X., and Nan, F. Discovery of 3-substituted 1 h-indole-2-carboxylic acid derivatives as a novel class of cyslt1 selective antagonists. *ACS Medicinal Chemistry Letters*, 7(3):335–339, 2016. 6
- Cho, K. pubmed-vectors: Dense Vector Retrieval for PubMed Abstracts. <https://github.com/kyunghyuncho/pubmed-vectors>, 2024. Accessed: 2026-01-19. 2
- Choung, O.-H., Vianello, R., Segler, M., Stiefl, N., and Jiménez-Luna, J. Extracting medicinal chemistry intuition via preference machine learning. *Nature Communications*, 14(1):6651, 2023. 1
- DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., et al. Deepseek-v3 technical report, 2025. URL <https://arxiv.org/abs/2412.19437>. 1, 3
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=zj7YuTE4t8>. 1, 8
- Du, Y., Yu, B., Liu, T., Shen, T., Chen, J., Rittig, J. G., Sun, K., Zhang, Y., Song, Z., Zhou, B., et al. Accelerating scientific discovery with autonomous goal-evolving agents. *arXiv preprint arXiv:2512.21782*, 2025. 1
- Gao, W., Fu, T., Sun, J., and Coley, C. W. Sample efficiency matters: A benchmark for practical molecular optimization. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL <https://openreview.net/forum?id=yCZRdI0Y7G>. 4, 24
- Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F.,

- Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., Matias, Y., Carroll, A., Kulkarni, K., Tomasev, N., Guan, Y., Dhillon, V., Vaishnav, E. D., Lee, B., Costa, T. R. D., Penadés, J. R., Peltz, G., Xu, Y., Pawlosky, A., Karthikesalingam, A., and Natarajan, V. Towards an ai co-scientist, 2025. URL <https://arxiv.org/abs/2502.18864>. 1
- Guo, J. and Schwaller, P. Augmented memory: sample-efficient generative molecular design with reinforcement learning. *Jacs Au*, 4(6):2160–2172, 2024. 5
- Jang, H., Jang, Y., Kim, J., and Ahn, S. Can llms generate diverse molecules? towards alignment with structural diversity. *arXiv preprint arXiv:2410.03138*, 2024. 4
- Jensen, J. H. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019. 5
- Kelder, J., Grootenhuys, P. D., Bayada, D. M., Delbressine, L. P., and Ploemen, J.-P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharmaceutical research*, 16(10):1514–1519, 1999. 6
- Kim, H., Kim, M., Choi, S., and Park, J. Genetic-guided gflownets for sample efficient molecular optimization. *Advances in Neural Information Processing Systems*, 37:42618–42648, 2024. 5
- Kim, H., Jang, Y., and Ahn, S. MT-mol: Multi agent system with tool-based reasoning for molecular optimization. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11544–11573, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.619. URL <https://aclanthology.org/2025.findings-emnlp.619/>. 1, 4, 5, 8
- Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., and Dong, X. Better zero-shot reasoning with role-play prompting. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4099–4113, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.228. URL <https://aclanthology.org/2024.naacl-long.228/>. 1, 2, 8
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019. 1
- Lee, S., Kreis, K., Veccham, S. P., Liu, M., Reidenbach, D., Peng, Y., Paliwal, S. G., Nie, W., and Vahdat, A. Genmol: A drug discovery generalist with discrete diffusion. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=KM7pXWG1xj>. 5
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 3
- Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL <https://aclanthology.org/2024.emnlp-main.992/>. 1, 8
- Liu, G., Zhao, H., Liu, B., Xin, Z., Liu, M., Kosogof, C., Szczepankiewicz, B. G., Wang, S., Clampit, J. E., Gum, R. J., et al. Aminopyridine carboxamides as c-jun n-terminal kinase inhibitors: targeting the gatekeeper residue and beyond. *Bioorganic & medicinal chemistry letters*, 16(22):5723–5730, 2006. 6
- Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL <https://arxiv.org/abs/2408.06292>. 1, 8
- Luna, D. pubmedFastRAG: Fast Retrieval-Augmented Generation for PubMed. <https://github.com/domluna/pubmedFastRAG>, 2024. Accessed: 2026-01-19. 2
- M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. 1, 8
- Mitchener, L., Yiu, A., Chang, B., Bourdenx, M., Nadolski, T., Sulovari, A., Landsness, E. C., Barabasi, D. L., Narayanan, S., Evans, N., Reddy, S., Foiani, M., Kamal, A., Shriver, L. P., Cao, F., Wassie, A. T., Laurent, J. M., Melville-Green, E., Caldas, M., Bou, A., Roberts, K. F., Zagorac, S., Orr, T. C., Orr, M. E., Zwezdaryk, K. J., Ghareeb, A. E., McCoy, L., Gomes, B., Ashley, E. A., Duff, K. E., Buonassisi, T., Rainforth,

- T., Bateman, R. J., Skarlinski, M., Rodrigues, S. G., Hinks, M. M., and White, A. D. Kosmos: An ai scientist for autonomous discovery, 2025. URL <https://arxiv.org/abs/2511.02824>. 8
- Nguyen, T. and Grover, A. LICO: Large language models for in-context molecular optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=yulvqQqKkx>. 4, 5, 8
- Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):48, 2017. 5
- Park, J. S., O’Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023. 1, 2, 8
- Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025. 3, 20
- Pedreira, J. G., Franco, L. S., and Barreiro, E. J. Chemical intuition in drug design and discovery. *Current topics in medicinal chemistry*, 19(19):1679–1693, 2019. 1
- Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., Lu, Z., Zheng, Z., Wang, J. Y., Zhou, D., et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025. 1, 8
- Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020. 3
- Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=Yacmpz84TH>. 1, 8
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023. 8
- Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., Nathan, A., Luo, A., Helyar, A., Madry, A., Efremov, A., Spyra, A., Baker-Whitcomb, A., Beutel, A., Karpenko, A., Makelov, A., Neitz, A., Wei, A., Barr, A., Kirchmeyer, A., et al. Openai gpt-5 system card, 2025. URL <https://arxiv.org/abs/2601.03267>. 1
- Srinivas, N., Krause, A., Kakade, S., and Seeger, M. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML’10*, pp. 1015–1022, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077. 5
- Su, H., Chen, R., Tang, S., Yin, Z., Zheng, X., Li, J., Qi, B., Wu, Q., Li, H., Ouyang, W., Torr, P., Zhou, B., and Dong, N. Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28201–28240, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1368. URL <https://aclanthology.org/2025.acl-long.1368/>. 1, 4, 5, 6, 8, 15
- Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: A family of highly capable multimodal models, 2025. URL <https://arxiv.org/abs/2312.11805>. 1
- Wang, H., Skreta, M., Ser, C. T., Gao, W., Kong, L., Strieth-Kalthoff, F., Duan, C., Zhuang, Y., Yu, Y., Zhu, Y., Du, Y., Aspuru-Guzik, A., Neklyudov, K., and Zhang, C. Efficient evolutionary search over chemical space with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=awWiNvQwf3>. 5, 8
- Wang, Z., Nie, W., Qiao, Z., Xiao, C., Baraniuk, R., and Anandkumar, A. Retrieval-based controllable molecule generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=vDFA1tpuLvK>. 5
- 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. 3
- Xie, Y., Xu, Z., Ma, J., and Mei, Q. How much space has been explored? measuring the chemical space covered by databases and machine-generated molecules. In *The*

Eleventh International Conference on Learning Representations, 2023. URL <https://openreview.net/forum?id=Yo06F8kfMa1>. 4

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X. 1, 8

Yi, P. and Qiu, M. 3d-qsar and docking studies of aminopyridine carboxamide inhibitors of c-jun n-terminal kinase-1. *European journal of medicinal chemistry*, 43(3):604–613, 2008. 6

Zhang, L., Yang, T., Xie, X., and Liu, G. Identification of 3, 5, 6-substituted indolin-2-one’s inhibitors of aurora b by development of a luminescent kinase assay. *Bioorganic & Medicinal Chemistry Letters*, 25(15):2937–2942, 2015. 6

Zheng, K., Iqbal, S., Hernandez, P., Park, H., LoGrasso, P. V., and Feng, Y. Design and synthesis of highly potent and isoform selective jnk3 inhibitors: Sar studies on aminopyrazole derivatives. *Journal of medicinal chemistry*, 57(23):10013–10030, 2014. 6

Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.-P., Bisk, Y., Fried, D., Neubig, G., and Sap, M. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=mM7VurbA4r>. 1, 2, 8

A. Prompts

In this section, we provide the full prompts including the system prompt, and task prompt used in our experiments.

A.1. INDIBATOR

Here, we provide the prompts employed for the agents in INDIBATOR. Below, we outline four prompts: (1) the system prompt used to initialize a scientist agent with a specific expertise profile, (2) the prompt that instructs the agents to suggest novel proposals with their scientific rationale, (3) the prompt that instructs the agents to evaluate proposals from their peers, and (4) the prompt that instructs the agents to score candidate molecules.

System prompt of scientist in INDIBATOR

You are {scientist_name}, a researcher specializing in molecular design and drug discovery.

Your expertise is based on your published work: {publications}
Molecules you have worked with: {molecules}

CURRENT TASK: {task_description}

In this debate, you will participate in three phases:

1. PROPOSAL PHASE: Propose 1-3 molecules (as valid SMILES strings) based on your expertise.
 - Draw from your knowledge of similar molecular scaffolds
 - Consider structure-activity relationships from your publications
 - Explain your rationale for each proposal

2. CRITIQUE PHASE: Evaluate other scientists' proposals.
 - Identify potential issues (toxicity, synthesis difficulty, selectivity)
 - Suggest modifications based on your experience
 - Note any overlap with molecules in your experience

3. VOTING PHASE: Score candidates from 0.0 to 1.0.
 - Consider task relevance, synthetic feasibility, and novelty
 - Base your assessment on your domain expertise

Always ground your contributions in your specific published expertise. When proposing molecules, provide valid SMILES strings and clear scientific rationale.

Prompt for proposal

Round {round_num} - PROPOSAL PHASE

Task: {task_description}

Previous proposals in this debate: {previous_proposals if previous_proposals else "No previous proposals yet."}

Based on your expertise, propose {num_mols}-{num_mols+2} novel molecules (as SMILES strings) that could address this task. For each molecule:

1. Provide the SMILES string
2. Explain your rationale based on your published work and proposed molecules in the published work
3. Discuss expected properties relevant to the task

Output format: [{ "SMILES": "SMILES string", "rationale": "Brief rationale" }, ...]

IMPORTANT:

- Output ONLY the JSON array, no other text or markdown
- Keep rationales brief (3-4 sentences) to avoid truncation
- Ensure all brackets and quotes are properly closed
- Do NOT wrap in code blocks
- Ensure the SMILES strings are valid and have not proposed in previous proposals
- Do NOT propose the duplicated molecules. Each proposal should be unique

Prompt for critique

Round {round_num} - CRITIQUE PHASE

Task: {task}

Review these proposals from other scientists: {proposals_to_critique}

Based on your expertise, critique each proposal:

- Identify potential issues considering the task description
- Suggest specific modifications if appropriate
- Note any overlap with molecules in your experience
- Highlight promising aspects that align with your research

Output format: [{ "SMILES": "SMILES string", "proposer": "Name", "critique": "Brief critique" }, ...]

IMPORTANT:

- Output ONLY the JSON array, no other text or markdown
- Keep critiques brief (3-4 sentences) to avoid truncation
- Ensure all brackets and quotes are properly closed
- Do NOT wrap in code blocks

Prompt for voting

Round {round_num} - VOTING PHASE

Task: {task}

All candidate molecules proposed in this debate: {all_candidates}

Vote for your top candidates by assigning scores from 0.0 to 1.0.

Format each vote as: SMILES: score

Consider:

- Relevance to the task
- Synthetic feasibility
- Novelty compared to existing drugs
- Critiques received from other scientists

Provide scores for at least top 3 candidates with brief justifications.

Output format: [{ "SMILES": "SMILES string", "score": 0.8, "justification": "Brief justification" }, ...]

IMPORTANT:

- Output ONLY the JSON array, no other text or markdown
- Keep justifications brief (3-4 sentences) to avoid truncation

- Score must be a number between 0.0 and 1.0
- Ensure all brackets and quotes are properly closed - Do NOT wrap in code blocks

A.2. Role persona

These prompts are employed for role persona baseline in Section 4.2, which defines agents based on role-based persona separation. Below, we outline two prompts: (1) the system prompt used to initialize a scientist agent with role-based persona, and (2) the prompt for the task-relevant role generation.

System prompt of role persona

You are {role_name}, an expert in molecular design and drug discovery.

Your expertise: {role_description}

CURRENT TASK: {task_description}

In this debate, you will participate in three phases:

1. PROPOSAL PHASE: Propose 1-3 molecules (as valid SMILES strings) based on your expertise.

- Draw from your knowledge of similar molecular scaffolds
- Consider structure-activity relationships from your publications
- Explain your rationale for each proposal

2. CRITIQUE PHASE: Evaluate other scientists' proposals.

- Identify potential issues (toxicity, synthesis difficulty, selectivity)
- Suggest modifications based on your experience
- Note any overlap with molecules in your experience

3. VOTING PHASE: Score candidates from 0.0 to 1.0.

- Consider task relevance, synthetic feasibility, and novelty
- Base your assessment on your domain expertise

Always ground your contributions in your specific expertise. When proposing molecules, provide valid SMILES strings and clear scientific rationale.

Prompt for role generation

Given the molecular optimization task: {task}

Generate {num_roles} distinct expert roles relevant to this task.

For each role, provide:

- role_name: A descriptive title (e.g., "Medicinal Chemist", "Computational Biologist", "Pharmacokinetics Expert")
- description: 3-4 sentences describing the expert's focus and how they contribute to the task

Output ONLY a valid JSON array, no other text:

[{"role_name": "...", "description": "..."}, ...]

A.3. Keyword persona

These prompts are employed for keyword persona baseline in Section 4.2, which is inspired by (Su et al., 2025). Below, we outline two prompts: (1) the system prompt used to initialize a scientist agent with keyword persona, and (2) the prompt for the research interest keyword extraction based on the publications.

System prompt of keyword persona

You are {scientist_name}, a researcher specializing in molecular design and drug discovery. You have researched on the following topics: keywords

CURRENT TASK: {task_description}

In this debate, you will participate in three phases:

1. PROPOSAL PHASE: Propose 1-3 molecules (as valid SMILES strings) based on your expertise.

- Draw from your knowledge of similar molecular scaffolds
- Consider structure-activity relationships from your publications
- Explain your rationale for each proposal

2. CRITIQUE PHASE: Evaluate other scientists' proposals.

- Identify potential issues (toxicity, synthesis difficulty, selectivity)
- Suggest modifications based on your experience
- Note any overlap with molecules in your experience

3. VOTING PHASE: Score candidates from 0.0 to 1.0.

- Consider task relevance, synthetic feasibility, and novelty
- Base your assessment on your domain expertise

Always ground your contributions in your specific published expertise. When proposing molecules, provide valid SMILES strings and clear scientific rationale.

Prompt for research interest keyword extraction

Extract the most important research keywords from the following publication abstracts.

Focus on:

- Research methodologies and techniques
- Target molecules, proteins, or biological systems
- Drug discovery concepts
- Chemical compound classes
- Disease areas or therapeutic targets

Publications: {publications}

Extract 10-20 unique, specific keywords that best characterize this researcher's expertise. Return ONLY a JSON array of keywords, no other text.

Example format: ["kinase inhibitors", "structure-activity relationship", "molecular docking", "EGFR", "cancer therapeutics"]

A.4. LLM-generated profile

These prompts are employed for LLM-generated profile baseline in Section 4.4. Below, we provide the prompts for LLM-based publication and molecule history profile that are relevant for the given task.

Prompt for LLM-based profile (publications) generation

You are tasked with generating a realistic publication summary for a researcher named {scientist_name} who specializes in molecular design and drug discovery.

Generate a list of 5-10 fictional but scientifically plausible publication titles that this researcher might have authored. Focus on publications relevant to: {task_description}

Output format (ONLY titles, one per line, each starting with "- "):

- Tetra-substituted pyridinylimidazoles as dual inhibitors of p38 α mitogen-activated protein kinase and c-Jun N-terminal kinase 3 for potential treatment of neurodegenerative diseases.
- 1,3-Dialkyl-substituted tetrahydropyrimido[1,2-f]purine-2,4-diones as multiple target drugs for the potential treatment of neurodegenerative diseases.

Make the publications diverse but coherent with the researcher's expertise in drug discovery and molecular optimization.

Output ONLY the list of titles in the exact format shown above.

Prompt for LLM-based profile (molecules) generation

You are tasked with generating a realistic molecule summary for a researcher named {scientist_name} who specializes in molecular design and drug discovery.

Generate a list of 5-10 molecules (as SMILES strings) that this researcher might have worked with in their publications. Focus on molecules relevant to: {task_description}

Output format (SMILES with publication title reference, ending with "..."):

- O=C(NCCN1CCOCC1)c1ccnc(-n2ncc3cc(Nc4ccccc4Cl)ccc32)c1 (from: Inhibitors of c-Jun N-terminal kinases: an update....)
- CNCC1CCCCC1(OC)c1cccc(OC)c1 (from: In vitro and in vivo evaluation of O-alkyl derivatives of tramadol....)

Ensure the SMILES strings are valid and represent drug-like molecules relevant to the task.

Output ONLY the list in the exact format shown above.

A.5. Vanilla

The prompt is employed for vanilla baseline in Section 3.2 and Section 3.3.

Prompt for vanilla generation

You are a researcher specializing in molecular design and drug discovery.

Task: {task_description}

Generate {batch_size} unique molecules.

Output format: [{ "SMILES": "SMILES string" }, { "SMILES": "SMILES string" }, ...]

Important:

- Output ONLY the JSON array, no other text or markdown
- Ensure all brackets and quotes are properly closed
- Do NOT wrap in code blocks
- Ensure the SMILES strings are valid and have not proposed in previous proposals
- Do NOT propose the duplicated molecules. Each proposal should be unique.

A.6. Task prompts

Protein target molecule generation The task prompt is employed for protein target molecule generation in Section 3.1.

Task prompt for protein target molecule generation

Your task is to design a SMILES string for a molecule that maximizes binding affinity to {protein_name}.

Objective: Design molecules with high predicted binding affinity (low IC50) to {protein_name}.

Evaluation Metrics: - affinity_pred_value: log10(IC50) in μM - LOWER values indicate STRONGER binding

Protein sequence: protein_sequence

Molecular Constraints: - Must be a valid SMILES string

Guidelines for High Binding Affinity:

- Include appropriate functional groups for hydrogen bonding (amines, hydroxyls, carbonyls)
- Consider hydrophobic contacts with protein binding pocket (aromatic rings, alkyl chains)
- Maintain reasonable molecular weight (300-600 Da)
- Include aromatic rings for pi-stacking interactions
- Consider salt bridges with charged residues (carboxylic acids, amines)

Generate ONLY the SMILES string with no explanation.

Bioactivity-guided molecule generation The task prompts are employed for bioactivity-guided molecule generation in Section 3.2 and Section 4.

Task prompt for GSK3 β bioactivity prediction

Your task is to design a SMILES string for a molecule that satisfies the following condition:

Conditions:

- Design a drug-like molecule with high predicted GSK3B (Glycogen Synthase Kinase 3 Beta) inhibitory activity.
- Consider structural features known to enhance kinase binding affinity.

IMPORTANT CONSTRAINTS:

- Design drug-like molecules with favorable ADMET properties.
- Maximize the GSK3B binding/inhibitory activity score as high as possible.
- Avoid generating identical structures to provided examples.
- Avoid repeating molecules you already generated.

Task prompt for DRD2 bioactivity prediction

Your task is to design a SMILES string for a molecule that satisfies the following condition:

Conditions:

Maximize the probability of binding to the DRD2 receptor (Dopamine Receptor D2).

IMPORTANT CONSTRAINTS:

- Design drug-like molecules.
- Maximize the DRD2 binding score as high as possible.
- Avoid generating identical structures to provided examples.

- Avoid repeating molecules you already generated.

Task prompt for JNK3 bioactivity prediction

Your task is to design a SMILES string for a molecule that satisfies the following condition:

Conditions:

- Design a drug-like molecule with high predicted JNK3 (c-Jun N-terminal kinase 3) inhibitory activity.
- Consider structural features known to enhance kinase inhibition.

IMPORTANT CONSTRAINTS:

- Design drug-like molecules with favorable ADMET properties.
- Maximize the JNK3 inhibitory activity score as high as possible.
- Avoid generating identical structures to provided examples.
- Avoid repeating molecules you already generated.

Goal-directed lead optimization The task prompt is employed for goal-directed lead optimization in Section 3.3.

Prompt for goal-directed lead optimization

Your task is to design a SMILES string for a molecule that optimizes binding affinity to {protein_name}.
You must start from the provided seed molecule, modify it to improve predicted binding affinity (docking score), while strictly satisfying all drug-likeness and similarity constraints.

Seed Molecule: {seed_mol}

Hard Constraints (MUST satisfy ALL):

1. Tanimoto Similarity \geq {sim_threshold}
2. QED (drug-likeness) \geq 0.6
3. SA Score (synthetic accessibility) \leq 4
4. NOT identical to seed molecule

Metric Definitions:

Tanimoto Similarity (target: \geq {sim_threshold})

Measures structural similarity between your molecule and the seed using Morgan fingerprints.

- Score range: 0 (completely different) to 1 (identical)
- Calculated as: (shared features) / (total unique features in both molecules)
- To maintain high similarity: preserve the core scaffold and make only small modifications

QED - Quantitative Estimate of Drug-likeness (target: \geq 0.6)

A composite score combining 8 drug-like properties, ranging 0-1 (higher = more drug-like).

Based on: molecular weight, logP, H-bond donors/acceptors, polar surface area, rotatable bonds, aromatic rings, and structural alerts.

To achieve QED \geq 0.6:

- Molecular Weight: 200-500 Da
- LogP: 0-5
- H-bond donors: \leq 5
- H-bond acceptors: \leq 10
- Rotatable bonds: \leq 10
- Aromatic rings: 1-4

SA Score - Synthetic Accessibility (target: ≤ 4)

Estimates how easy a molecule is to synthesize, ranging 1 (easy) to 10 (very difficult).

Calculated from fragment contributions (based on 1M PubChem molecules) plus complexity penalties for unusual features.

To achieve $SA \leq 4$:

- Use common, commercially available building blocks
- Avoid: large rings (≥ 8 atoms), bridgehead/spiro atoms, multiple stereocenters
- Prefer simple ring systems (benzene, pyridine, piperidine)

Recommended Modifications (preserve similarity, maintain drug-likeness):

- Small substituent changes: $-H \rightarrow -F$, $-CH_3 \rightarrow -CF_3$, $-OH \rightarrow -OCH_3$
- Bioisosteric replacements: benzene \leftrightarrow pyridine, $-COOH \leftrightarrow -CONH_2$
- Methylation of amines: $-NH_2 \rightarrow -NHCH_3$
- Small ring modifications: 6-ring \rightarrow 5-ring

Modifications to AVOID:

- Adding large polycyclic systems (hurts SA)
- Adding ≥ 2 new rings (may hurt similarity)
- Adding unusual functional groups (hurts SA)
- Removing core scaffold elements (hurts similarity)

Generate ONLY the SMILES string with no explanation.

B. Experimental settings

We provide the code in https://anonymous.4open.science/r/debate_scientist-AACB.

B.1. Hyperparameters

We configure the collaboration size to $N = 50$ scientist agents. In the proposal phase, each agent generates $k = 30$ candidate molecules per iteration. To guarantee that the debate yields a sufficient volume of molecular candidates, we set the maximum number of rounds to 20. For all model sampling, we utilize a temperature of 0.7.

B.2. Computational resource

We utilized a single NVIDIA RTX A5000 GPU for Boltz-2 (Passaro et al., 2025) prediction for protein target molecule generation task in Section 3.1.

C. Additional experimental results

C.1. Detailed examples

Here, we provide the detailed examples of debate process in Figure 3.

A detailed example of publication history.

```
{
  "title": "3D-QSAR and docking studies of aminopyridine carboxamide inhibitors of c-Jun N-terminal kinase-1.",
  "abstract": "In order to better understand the structural and chemical features of c-Jun N-terminal kinase-1 (JNK-1), which is a member of the mitogen activated protein kinase (MAP kinase) family of enzymes responsible for the serine/threonine phosphorylation of intracellular targets, 3D-QSAR studies of some aminopyridine carboxamides as c-Jun N-terminal kinase inhibitors were performed by comparative molecular field analysis (CoMFA) to rationalize the structural requirements responsible for the inhibitory activity of these compounds. The genetic algorithm of GOLD3.1 has been employed to position 54 aminopyridine carboxamides in the active sites of JNK-1 to determine the probable binding conformation. The docking results provided a reliable conformational alignment scheme for 3D-QSAR model. Based on the docking conformations, highly predictive comparative molecular field analysis (CoMFA) was performed with a cross-validated q(2) of 0.585. The non-cross-validated analysis with six optimum components revealed a conventional r(2) value of 0.988, F=510.200, and an estimated standard error of 0.071. Furthermore, the CoMFA model was mapped back to the binding sites of JNK-1, to get a better understanding of vital interactions between the aminopyridine carboxamides and the kinase. Based on the docking and CoMFA analyses, we have identified some key features in the aminopyridine carboxamides that are responsible for JNK-1 inhibitory activity. The analyses may be used to design more potent aminopyridine carboxamides and predict their activity prior to synthesis."
},
{
  "title": "Integrating Metabolomics and Proteomics Technologies Provides Insights into the Flavor Precursor Changes at Different Maturity Stages of Arabica Coffee Cherries.",
  "abstract": "The metabolic modulation of major flavor precursors during coffee cherry ripening is critical for the characteristic coffee flavor formation. However, the formation mechanism of flavor precursors during coffee cherry ripening remains unknown. In the present study, a colorimeter was employed to distinguish different maturity stages of coffee cherry based on the coffee cherry skin colors, and proteomics and metabolomics profiles were integrated to comprehensively investigate the flavor precursor dynamics involved in Arabica coffee cherry ripening. The data obtained in the present study provide an integral view of the critical pathways involved in flavor precursor changes during coffee cherry ripening. Moreover, the contributions of critical events in regulating the development of flavor precursors during the four ripening stages of coffee cherries, including the biosynthesis and metabolism pathways of organic acids, amino acids, flavonoids, and sugars, are discussed. Overall, a total of 456 difference express metabolites were selected, and they were identified as being concentrated in the four maturity stages of coffee cherries; furthermore, 76 crucial enzymes from the biosynthesis and metabolism of sugars, organic acids, amino acids, and flavonoids contributed to flavor precursor formation. Among these enzymes, 45 difference express proteins that could regulate 40 primary amino acids and organic acids flavor precursors were confirmed. This confirmation indicates that the metabolic pathways of amino acids and organic acids played a significant role in the flavor formation of Arabica coffee cherries during ripening. These results provide new insights into the protease modulation of flavor precursor changes in Arabica coffee cherry ripening."
},
}
```

A detailed example of molecule history.

```
{
  "smiles": "O=C1NC(=O)c2c1c(-c1ccccc1Cl)cc1oc3ccc(O)cc3c21",
  "inchi_key": "KGAPFKLOTLAAMY-UHFFFAOYSA-N",
  "title": "3D-QSAR studies of Checkpoint Kinase Weel inhibitors based on molecular docking, CoMFA and
```

```

CoMSIA.",
"pubmed_id": "17698256"
},
{
"smiles": "CCn1c2ccc(O)cc2c2c3c(c(-c4ccccc4Cl)cc21)C(=O)NC3=O",
"inchi_key": "PPKWYAYCBJKWOR-UHFFFAOYSA-N",
"title": "3D-QSAR studies of Checkpoint Kinase Weel inhibitors based on molecular docking, CoMFA and CoMSIA.",
"pubmed_id": "17698256"
}

```

A detailed example of proposal phase.

```

{
"SMILES": "CS(=O)(=O)c1ccc(CNC(=O)c2cc(N)c(C#N)c(NC3CCCC3)n2)cc1",
"rationale": "This molecule is designed based on structural insights from my paper '3D-QSAR and docking studies of aminopyridine carboxamide inhibitors of c-Jun N-terminal kinase-1.' The core scaffold features an aminopyridine carboxamide, which is a known pharmacophore for JNK inhibition. The sulfonamide group (CS(=O)(=O)) at the para position of the phenyl ring, as seen in several active compounds from that study, is crucial for binding affinity and selectivity. The cyano group (C#N) adjacent to the pyridine nitrogen enhances electron-withdrawing properties, potentially improving kinase hinge region interaction. The cycloalkylamine substituent (NC3CCCC3) on the pyridine ring is introduced to modulate lipophilicity and metabolic stability, aiming for favorable ADMET properties. This design integrates key features from successful JNK1 inhibitors, adapted here for JNK3, which shares high structural homology in the ATP-binding site."
},
{
"SMILES": "CC(=O)Nc1ccc(CNC(=O)c2cc(N)c(C#N)c(NC3CC3)n2)cc1",
"rationale": "An acetamide group (CC(=O)N) replaces the sulfonamide, exploring alternative hydrogen bond acceptor/donor profiles for JNK3 binding, based on the versatility of amides in kinase inhibitors noted in my QSAR studies. The core cyanoaminopyridine with cyclopropylamine is maintained. The acetyl group may be metabolically labile, offering a potential prodrug effect. This design diversifies from the typical sulfonamide, aiming to discover new interactions while retaining the essential pharmacophore. It reflects the structural variation seen in my work on different triterpene classes, such as cycloartanes versus serratanes, where subtle changes led to distinct biological activities."
}

```

A detailed example of critique phase.

```

{
"SMILES": "CS(=O)(=O)c1ccc(CNC(=O)c2cc(N)c(C#N)c(NC3CCCC3)n2)cc1",
"proposer": "XXX XXX",
"critique": "Cyclopentylamine is a good hydrophobic group. The sulfonamide remains a concern for metabolism. The molecule is similar to others proposed; consider varying the core scaffold to explore different binding modes, perhaps by incorporating a heterocyclic hinge binder like an indole or purine."
},
{
"SMILES": "CS(=O)(=O)c1ccc(CNC(=O)c2cc(N)c(C#N)c(NC3CCCC3)n2)cc1",
"proposer": "XXX XXX",
"critique": "The cyclobutyl group (CCCC3) is a constrained aliphatic ring, which might improve potency by reducing flexibility. However, the hinge-binding motif is still not optimal. From my research (Papers 1, 6), kinase inhibitors benefit from specific heterocyclic cores."
}

```

A detailed example of voting phase.

```
{
  "SMILES": "CS(=O)(=O)c1ccc(CNC(=O)c2cc(N)c(C#N)c(NC3CCCC3)n2)cc1",
  "score": 0.9,
  "justification": "Based on my 3D-QSAR studies of aminopyridine carboxamide JNK1 inhibitors (Paper 2), this scaffold shows high predicted JNK3 affinity. The sulfonylphenyl linker and cyanoaminopyridine core are key for kinase binding. The cyclobutylamide tail enhances selectivity and ADMET properties, as seen in my previous work."
},
{
  "SMILES": "O=c1ncn2nc(Sc3ccccc3)ccc2c1-c1ccc(C(F)(F)F)cc1",
  "score": 0.85,
  "justification": "This design incorporates a trifluoromethyl group, known to improve lipophilicity and target engagement in kinase inhibitors. My research on antiepileptic drugs (Paper 1,3,6,8,9) highlights the importance of halogenated aromatics for CNS penetration, which is relevant for JNK3 inhibition in neurological contexts. The scaffold is synthetically accessible and avoids toxicity concerns from heavy halogens."
}
```

C.2. Number of scientists and proposals

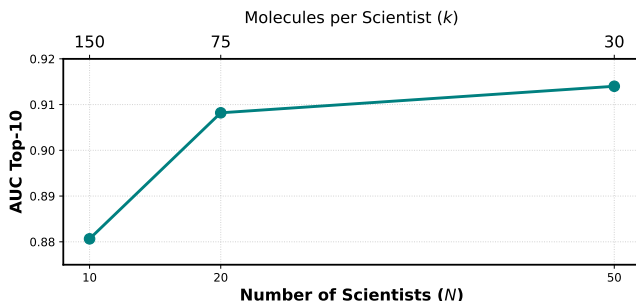


Figure 5. Fixed total number of proposals.

To analyze the impact of the number of scientists and the number of proposals per scientist, we evaluate the interplay between the number of scientists that debate (N) and the number of proposals that one scientist proposes per round (k) in JNK bioactivity optimization task.

First, we fix the total number of proposals $N \times k = 1,500$. This allows us to investigate the trade-off between scientist diversity (increasing N) against exploration depth (increasing k). By varying the composition, we observe whether the scientist diversity is more critical than the depth of expertise.

We provide the results in Figure 4. This demonstrates demonstrate that increasing the number of scientists (N) consistently enhances discovery performance across both experimental settings. In the fixed total budget scenario, the system maintains relatively high AUC scores even at low N , suggesting that increased exploration depth (k) can partially compensate for limited expertise diversity.

C.3. Detailed protein-conditioned molecule generation results

Table 5. Results of protein target molecule generation (binding affinity). Bold highlights the best scores.

	TYK2		JNK1		CDK2		P38		CA2		DHFR		FABP4		THROMBIN	
	Top1	Top10	Top1	Top10	Top1	Top10	Top1	Top10	Top1	Top10	Top1	Top10	Top1	Top10	Top1	Top10
VANILLADEBATE	-9.92	-9.48	-8.99	-8.75	-10.00	-9.20	-9.42	-9.04	-11.09	-10.92	-8.68	-8.17	-8.40	-7.87	-9.35	-8.70
KEYWORDDEBATE	-9.73	-9.40	-9.49	-9.13	-9.48	-9.03	-9.77	-9.08	-11.00	-10.62	-9.22	-8.86	-9.00	-8.13	-9.73	-8.50
INDIBATOR (Ours)	-11.97	-10.71	-10.51	-10.08	-10.52	-10.12	-10.88	-10.37	-12.76	-11.68	-11.36	-10.12	-9.65	-9.23	-11.33	-10.71

Table 6. Results of protein target molecule generation (diversity). Bold highlights the best scores.

	TYK2			JNK1			CDK2			P38		
	IntDiv	#Cir _{0.75}	#Cir _{0.85}	IntDiv	#Cir _{0.75}	#Cir _{0.85}	IntDiv	#Cir _{0.75}	#Cir _{0.85}	IntDiv	#Cir _{0.75}	#Cir _{0.85}
VANILLADEBATE	0.8079	53	9	0.7957	40	7	0.8021	47	9	0.8085	55	9
KEYWORDDEBATE	0.7941	42	7	0.8103	49	8	0.8034	41	6	0.8096	45	8
INDIBATOR (Ours)	0.8695	203	35	0.8691	229	38	0.8661	211	45	0.8699	239	43

	CA2			DHFR			FABP4			THROMBIN		
	IntDiv	#Cir _{0.75}	#Cir _{0.85}	IntDiv	#Cir _{0.75}	#Cir _{0.85}	IntDiv	#Cir _{0.75}	#Cir _{0.85}	IntDiv	#Cir _{0.75}	#Cir _{0.85}
VANILLADEBATE	0.7468	20	6	0.7821	37	9	0.7529	35	7	0.7408	27	5
KEYWORDDEBATE	0.7274	23	6	0.7990	42	8	0.7130	27	7	0.7539	26	5
INDIBATOR (Ours)	0.8655	139	46	0.8679	192	40	0.8714	250	44	0.8576	165	36

Here, we provide detailed protein conditioned molecule generation results in Table 5.

C.4. Additional PMO tasks

Table 7. Results of PMO-1K benchmark. Tasks are assessed using top-10 AUC. We mark the best result in bold and teal highlights the improvement to the vanilla-debate.

Model	Bioactivity			Multi property optimization							Rediscovery		
	GSK3 β	DRD2	JNK3	Amlo.	Fexo.	Osim.	Peri.	Rano.	Sita.	Zale.	Cele.	Thio.	Trog.
GP BO	0.611	0.857	0.346	0.519	0.707	0.766	0.458	0.701	0.232	0.392	0.411	0.351	0.313
REINVENT	0.589	0.775	0.315	0.472	0.650	0.737	0.404	0.574	0.261	0.406	0.370	0.311	0.246
LICO-L	0.617	0.859	0.336	0.541	0.700	0.759	0.473	0.687	0.315	0.404	0.447	0.343	0.292
Genetic GFN	0.637	0.809	0.409	0.534	0.682	0.763	0.462	0.623	0.227	0.400	0.447	0.377	0.277
Graph GA	0.523	0.833	0.301	0.501	0.666	0.751	0.435	0.620	0.229	0.374	0.424	0.322	0.267
Aug. Mem.	0.539	0.795	0.294	0.489	0.679	0.761	0.422	0.614	0.245	0.415	0.385	0.336	0.262
MOLLEO-B	0.397	0.910	0.186	0.637	0.674	0.779	0.655	0.640	0.193	0.392	0.402	0.416	0.302
MOLLEO-D	0.496	0.812	0.342	0.540	0.680	0.753	0.422	0.516	0.328	0.409	0.512	0.478	0.387
MT-Mol	0.308	0.756	0.125	0.647	0.883	0.796	0.542	0.233	0.067	0.625	0.867	0.719	0.841
Vanilla	0.419	0.921	0.310	0.854	0.530	0.877	0.678	0.642	0.099	0.708	0.825	0.845	0.825
Vanilla-debate	0.477	0.902	0.161	0.856	0.935	0.939	0.769	0.636	0.310	0.740	0.819	0.808	0.820
INDIBATOR (Ours)	0.942	0.950	0.914	0.845	0.925	0.941	0.775	0.848	0.225	0.730	0.821	0.831	0.838

For completeness, we provide additional results on the PMO benchmark (Gao et al., 2022), including multi-property optimization and molecule rediscovery tasks. Although these tasks were excluded from the main text due to their nature as arithmetic structural puzzles, we report their performance in Table 7. We observe that INDIBATOR still achieves consistent performance improvements in most cases compared to the vanilla debate. However, we emphasize that these metrics are less indicative of our framework’s true utility, as they do not require the individual profile-grounded reasoning or the broad chemical space exploration that INDIBATOR is designed to facilitate.