# LIDDIA: Language-based Intelligent Drug Discovery Agent

**Reza Averly[*1], Frazier N. Baker[*1], Ian A. Watson[2] & Xia Ning[1,3,4,5]**
[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Independent Researcher
[3]Department of Biomedical Informatics, The Ohio State University, USA
[4]Translational Data Analytics Institute, The Ohio State University, USA
[5]College of Pharmacy, The Ohio State University, USA
[*]Equal contribution
{averly.1,baker.3239}@buckeyemail.osu.edu,ianiwatson@gmail.com,
ning.104@osu.edu

## Abstract

Drug discovery is a long, expensive, and complex process, relying heavily on human medicinal chemists, who can spend years searching the vast space of potential therapies. Recent advances in artificial intelligence for chemistry have sought to expedite individual drug discovery tasks; however, there remains a critical need for an intelligent agent that can navigate the drug discovery process. Towards this end, we introduce LIDDIA, an autonomous agent capable of intelligently navigating the drug discovery process *in silico*. By leveraging the reasoning capabilities of large language models, LIDDIA serves as a low-cost and highly-adaptable tool for autonomous drug discovery. We comprehensively examine LIDDIA, demonstrating that (1) it can generate molecules meeting key pharmaceutical criteria on over 70% of 30 clinically relevant targets, (2) it intelligently balances exploration and exploitation in the chemical space, and (3) it identifies one promising novel candidate on AR/NR3C4, a critical target for both prostate and breast cancers. Code and dataset are available at https://github.com/ninglab/LIDDiA.

## 1 Introduction

Artificial intelligence (AI) research has long sought to develop agents capable of intelligent reasoning to aid humans by autonomously navigating complex, resource-intensive processes. Drug discovery is one such process, relying heavily on human medicinal chemists, who can spend years searching the vast space of potential therapies (Blass, 2021). Recent advances (Chen et al., 2020; Zhao et al., 2024; Trott and Olson, 2010; Swanson et al., 2024; Zhou et al., 2019; Jensen, 2019) in AI for chemistry have sought to expedite drug discovery by performing individual tasks *in silico*. However, there remains a critical need for an intelligent, autonomous agent that can strategically navigate and facilitate the drug discovery process.

Drug discovery is a complex, nonlinear process with many requirements. Successful drugs must not only bind well to their therapeutic targets, but also exhibit good physicochemical, pharmacodynamic, and pharmacokinetic properties. These requirements are not necessarily independent; changing a molecule to satisfy one may result in the violation of another. Medicinal chemists combine manual analysis with computational tools to identify promising molecules, evaluate their properties, and optimize their structures—an iterative process that demands substantial time and effort.

Large language models (LLMs) have emerged as reasoning engines capable of intelligent reasoning and planning over complex tasks. Recent works (Yao et al., 2022; Liu et al., 2023; Boiko et al., 2023; Zhou et al., 2023) have explored leveraging LLMs as intelligent agents, using natural language as an interface for taking actions and observing results. By pairing LLM's reasoning capabilities with computational tools for drug discovery, we envision building a digital twin of the medicinal chemist, capable of navigating the complexities of the drug discovery process.

In this work, we introduce LIDDIA, an intelligent agent for navigating the pre-clinical drug discovery process *in silico*. LIDDIA is composed of four interconnected components: (1) REASONER, (2) EXECUTOR, (3) EVALUATOR, and (4) MEMORY. Each component interacts with the others to collaboratively navigate the drug discovery process. By harnessing the pre-trained knowledge and reasoning capabilities of LLMs, LIDDIA enables intelligent and rational decision-making over drug discovery steps, mimicking experienced medicinal chemists and steering the drug discovery process toward high throughput and success rate. In doing so, LIDDIA orchestrates the intelligent use of computational tools (e.g., docking simulation, property prediction, molecule optimiza-



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12004–12028
November 4-9, 2025 

tion). One key strength of LIDDIA lies in its integration of generative AI tools for molecular design, enabling it to explore vast chemical spaces beyond conventional molecular libraries. With a modular architecture, LIDDIA is designed for flexibility, allowing it to be seamlessly extended or refined as new capabilities emerge. To the best of our knowledge, LIDDIA is the first of its kind, representing the first effort toward low-cost, high-efficiency, autonomous drug discovery.

We rigorously benchmark LIDDIA (Section 5) and demonstrate that it can produce promising drug candidates satisfying key pharmaceutical properties on more than 70% of 30 major therapeutic targets (Section 5.1). We provide an in-depth study (Section 5.2), illustrating that LIDDIA strategically generates, refines, and selects highly favorable molecules, well aligned with a real-world drug discovery workflow (Section 5.2.1). We also identify a salient pattern underpinning successful outcomes for LIDDIA: effectively balancing exploration and exploitation in the chemical space (Section 5.2.2). Lastly, we highlight one promising drug candidate for AR/NR3C4, an important target for prostate and breast cancers (Section 5.4).

## 2 Related Work

LLMs equipped with tools have recently shown great promise as autonomous agents for scientific discoveries, including drug discovery (Gao et al., 2024). For instance, AutoBA (Zhou et al., 2023) uses LLMs to automate multi-omics bioinformatics analysis, generating new insights using well-known bioinformatics libraries. Another example for biomedical research is PROTAGENT (Ghafarollahi and J. Buehler, 2024), an LLM agent system for *de novo* protein design, equipped with physics-based simulations to ground its design process. Materials science can also benefit from LLM agents as well, as shown by A-LAB (Szymanski et al., 2023), a self-driving laboratory that uses an LLM agent to control both analysis tools and laboratory hardware for semiconductor material design.

For drug discovery, COSCIENTIST (Boiko et al., 2023) uses LLMs to perform web search, conduct technical documentation, program, and operate physical hardware modules to plan and control chemical synthetic experiments. It demonstrates the viability LLM agents equipped with both physical and computational tools to act as self-driving laboratories for organic chemistry. However, COSCIENTIST does not integrate any domain-specific tools for grounding, but rather relies upon the LLM's intrinsic chemistry knowledge, web search, and results from the physical experiments.

In addition, CHEMCROW (M. Bran et al., 2024) is an LLM agent equipped with specific tools for small molecule organic chemistry. These tools support generating molecule structures from natural language descriptions, predicting molecule properties, conducting *in silico* safety checks, and performing retrosynthesis planning. Grounded by these tools, CHEMCROW demonstrates an ability to perform complex, multi-step chemistry tasks commonly found in the drug discovery process. CACTUS (McNaughton et al., 2024) is a similar agent to CHEMCROW, emphasizing tools that can predict properties important to drug discovery. DRUGAGENT (Inoue et al., 2024) is an LLM agent for drug repurposing, equipped with tools to search databases of existing drugs to identify candidates likely to interact with a protein target. Notably, none of these LLM agents are grounded with well-established computational tools for novel structure-based drug discovery (SBDD).

## 3 LIDDIA Framework

LIDDIA is an automated, agentic framework for navigating the drug discovery process by combining computational tools and reasoning capabilities. As illustrated in Figure 1, LIDDIA is composed of four components: (1) REASONER (Section 3.1), which plans LIDDIA's actions and directs LIDDIA to conduct drug discovery; (2) EXECUTOR (Section 3.2), which executes REASONER's actions using state-of-the-art computational tools; (3) EVALUATOR (Section 3.3), which assesses candidate molecules; and (4) MEMORY (Section 3.4), which keeps all the information produced along the drug discovery process. Each of these components represents a logical abstraction from the traditional drug discovery, enhanced by computational tools and generative AI. The ultimate goal is, given a target of interest and property specifications on its potential drugs (e.g., at least 5 molecules, binding affinities better than -7, drug-likeness better than 0.5), LIDDIA produces a diverse set of high-quality molecules that satisfy these specifications and can be considered as potential drug candidates for the target. Overall, LIDDIA represents an innovative initiative towards autonomous drug discovery, integrat-

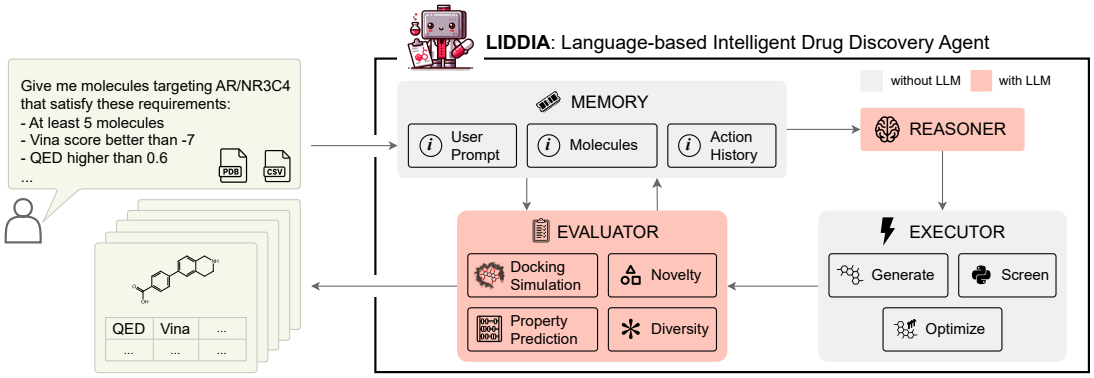



Figure 1: Overview of the LIDDIA Framework

ing AI-driven planning, execution, evaluation, and memory management to accelerate the identification and optimization of novel therapeutics.

### 3.1 REASONER

LIDDIA's decision-making is conducted through its REASONER component. Using the information in MEMORY (e.g., molecules under current consideration and their property profiles), REASONER conducts reasoning and strategically plans the next actions that LIDDIA should take, leveraging the pre-trained knowledge and reasoning capabilities of LLMs. REASONER explores three action types: (1) GENERATE to generate new molecules; (2) OPTIMIZE to optimize existing molecules; and (3) SCREEN to process the current molecules. These actions correspond to several key steps in hit identification (via generative AI) and lead optimization in pre-clinical drug discovery. Therefore, REASONER is key to guiding LIDDIA through the iterative process of molecular design, ensuring that each decision aligns with the overall requirements of drug discovery while balancing all required properties.

### 3.2 EXECUTOR

LIDDIA executes all drug discovery actions planned by REASONER through its EXECUTOR. EXECUTOR is equipped with state-of-the-art computational tools tailored to different actions. Specifically, EXECUTOR integrates: (1). generative models for structure-based drug design to implement the GENERATE action, utilizing methods such as Pocket2Mol (Peng et al., 2022); (2). generative models for molecular refinement to implement the OPTIMIZE action, enhancing drug-like properties and optimizing molecular structures (Jensen, 2019); and (3). a processor to implement the SCREEN action, allowing complex and logic screening, organizing and managing molecules, and identifying the most promising ones.

A key innovation of LIDDIA is that EXECUTOR leverages generative models for both hit identification (GENERATE) and lead optimization (OPTIMIZE). Unlike conventional drug discovery, which relies on searching and modifying existing molecular databases, EXECUTOR enables *de novo* molecular generation, expanding the chemical space beyond known molecules. This approach increases the likelihood of discovering novel, diverse, and more effective drug candidates. Moreover, by automating lead optimization through generative models, EXECUTOR reduces human bias stemming from individual expertise levels and limited chemical knowledge. This ensures a more systematic, data-driven approach to improving molecular properties. By integrating generative tools, LIDDIA can operate autonomously, designing superior drug candidates more efficiently than traditional human-driven search methods. This automation enhances cost-effectiveness and accelerates drug discovery, making LIDDIA a powerful AI-driven co-pilot in the drug discovery process.

### 3.3 EVALUATOR

LIDDIA performs *in silico* assessments over molecules using its EVALUATOR. EVALUATOR assesses an array of molecule properties essential to successful drug candidates, including target binding affinity, drug-likeness, synthetic accessibility, Lipinski's rule, novelty, and diversity. For different properties, EVALUATOR uses the appropriate computational tools to conduct the evaluation. Evaluation results are systematically stored in LIDDIA's memory (MEMORY; discussed in Section 3.4), and subsequently utilized by REASONER to refine decision-making and guide the next steps in the drug discovery process. Once EVALUATOR identifies molecules satisfying all user requirements, it signals LIDDIA to terminate the search and return

Table 1: Statistics over protein targets.

| Disease | #Targets (%) |
|---|---|
| Cancers | 15 (50%) |
| Neurological Conditions | 8 (27%) |
| Cardiovascular Diseases | 6 (20%) |
| Infectious Diseases | 4 (13%) |
| Diabetes | 3 (10%) |
| Autoimmune Diseases | 3 (10%) |

Some targets are associated with multiple categories.

the most promising candidates. Thus, EVALUATOR serves as a critical quality control mechanism, systematically steering LIDDIA toward identifying optimal drug candidates while minimizing the exploration of suboptimal chemical spaces. In addition, EVALUATOR is designed in a plug-and-play fashion, allowing additional tools to be added or updated to support new property requirements.

### 3.4 MEMORY

LIDDIA keeps all the information produced throughout its entire drug discovery process in MEMORY. This includes information provided by users via prompts, such as protein target structures, property requirements, and reference molecules (e.g., known drugs). More information will be dynamically generated as LIDDIA progresses through the drug discovery process, including the trajectories of actions that LIDDIA has taken (planned by REASONER), molecules generated from prior actions, and their properties. The information in MEMORY is aggregated and provided to REASONER to facilitate its planning. MEMORY is dynamically changing and continuously updated. This evolving repository enables REASONER to be well informed by prior knowledge and newly generated data, and thus better reflect and refine its strategies, enhancing the efficiency and effectiveness of automated drug discovery.

## 4 Experimental Settings

### 4.1 Evaluation Metrics

**Molecule Qualities** We use these metrics to evaluate the molecules generated by different methods.

**Key molecule properties** We first evaluate the following general properties required for successful drugs: **(1)** drug-likeness (Bickerton et al., 2012) (QED), **(2)** Lipinski's Rule of Five (Lipinski et al., 2001) (LRF), **(3)** synthetic accessibility (Ertl and Schuffenhauer, 2009) (SAS), and **(4)** binding affinities measured by Vina scores (Trott and Olson, 2010) (VNA). Evaluation on more molecule properties (e.g. toxicity properties) is available in Appendix C.2.

**Novelty** We measure the novelty (NVT) of a molecule $m$ with respect to a reference set of known drugs $\mathcal{M}_0$ as follows:

$$\texttt{NVT}(m; \mathcal{M}_0) = 1 - \max_{m_i \in \mathcal{M}_0}(\text{sim}_\intercal(m, m_i)),$$

where $\mathcal{M}_0$ is the reference set of known drugs, $m$ and $m_i$ are two molecules, and $\text{sim}_\intercal(m, m_i)$ is the Tanimoto similarity of $m$ and $m_i$'s Morgan fingerprints (Morgan, 1965). High novelty indicates that new molecules are different from existing drugs, offering new therapeutic opportunities. A molecule $m$ is considered novel if $\texttt{NVT}(m) \geq 0.8$.

**High-quality molecules** A molecule $m$ is considered as "high quality" (HQ) for a target $t$, if its properties satisfy $\texttt{QED} \geq \overline{\texttt{QED}}_t$, $\texttt{LRF} \geq \overline{\texttt{LRF}}_t$, $\texttt{SAS} \leq \overline{\texttt{SAS}}_t$, $\texttt{VNA} \leq \overline{\texttt{VNA}}_t$, and $\texttt{NVT}(m) \geq 0.8$, where the overline and the subscript $t$ indicate the average value from all the known drugs for target $t$. Such multi-property requirements are typical in drug discovery. Meanwhile, this presents a significant challenge, as LIDDIA must identify molecules with key properties similar to or even better than existing drugs but structurally significantly different from them. In our dataset, we include existing drugs for targets as the gold standard for evaluation purposes.

**Molecule Set Diversity** We measure the diversity (DVS) of a set of generated molecules $\mathcal{M}$ defined as follows,

$$\texttt{DVS}(\mathcal{M}) = 1 - \mathbb{E}_{\{m_i, m_j\} \subseteq \mathcal{M}}\left[\text{sim}_\intercal(m_i, m_j)\right],$$

where $m_i$ and $m_j$ are two distinct molecules in $\mathcal{M}$. High diversity is preferred, as chemically diverse molecules increase the likelihood of identifying successful drug candidates. A set of molecules $\mathcal{M}$ is considered diverse if $\texttt{DVS}(\mathcal{M}) \geq 0.8$. This imposes a highly stringent requirement on the diversity of the generated molecules.

**Target Success Rate** Target success rate, denoted as TSR, is defined as the percentage of targets for which a method can generate a *diverse* set of *at least 5 high-quality* molecules.

### 4.2 Protein Target Dataset

To evaluate LIDDIA, we manually curated a diverse set of protein targets from OpenTargets (Ochoa et al., 2023) that are strongly associated with major human diseases: cancers, neurological conditions, cardiovascular diseases, in-

fectious diseases, diabetes, and autoimmune diseases. For each of these protein targets, we identified an experimentally resolved structure with a small-molecule ligand from the RCSB Protein Data Bank (PDB) (Berman et al., 2000) and extracted the binding pocket according to its ligand's position. To enable a comparison to existing drugs, we searched ChEMBL (Bento et al., 2014; Gaulton et al., 2011) for all known drugs targeting the selected proteins. This leads to 30 protein targets with PDB structures, ligands, and existing drugs. These targets will be used as input in our experiments. Table 1 presents the distribution of the targets in terms of their disease associations. Please note, some targets are associated with multiple diseases. A full list of targets is presented in Appendix A. We discuss the importance of manual curation of this dataset in Appendix A.1.

### 4.3 Implementation Details

LIDDIA leverages Claude 3.5 Sonnet (Anthropic, 2024) as the base model for its REASONER and EVALUATOR since it achieves state-of-the-art performance in chemistry related tasks (Chen et al., 2024; Huang et al., 2024). We designed and fine-tuned specific prompts to guide REASONER and EVALUATOR, respectively. Details on these prompts are provided in Appendix B. EVALUATOR evaluates all the metrics as defined in Section 4.1. We set the maximum number of actions taken to 10 to ensure a concise yet effective drug discovery trajectory.

EXECUTOR executes the GENERATE action using Pocket2Mol (Peng et al., 2022). As a structure-based drug design tool, Pocket2Mol can generate molecules using only the target protein structure. This provides LIDDIA with the ability to extend to novel targets without known ligands. For efficiency, Pocket2Mol is set to generate a minimum of 100 molecules using a beam size of 300. The OPTIMIZE action is implemented via GraphGA (Jensen, 2019), a popular graph-based genetic algorithm for molecule optimization. In LIDDIA, OPTIMIZE can refine molecules on three essential properties: drug-likeness (QED), synthetic accessibility (SAS), and target binding affinity (VNA). However, the actions can be easily expanded to cover additional properties.

### 4.4 Baselines

We compare LIDDIA with two types of baselines: task-specific molecule generation methods, and general-purpose LLMs. For molecule generation methods, we use Pocket2Mol (Peng et al., 2022) and DiffSMol (Chen et al., 2025). Pocket2Mol (Peng et al., 2022) is a well-established generative method for structure-based drug design, which uses binding pocket structures as input. DiffSMol (Chen et al., 2025), on the other hand, is a state-of-the-art generative method for ligand-based drug design, requiring a binding ligand. These two methods represent distinct approaches in computational drug design, using different information to generate potential drug candidates. Notably, Pocket2Mol is used by LIDDIA in GENERATE actions, capitalizing on the popularity of SBDD and its ability to generate molecules without reference ligands.

For general-purpose LLMs, we use GPT-4o (OpenAI et al., 2024), o1 (OpenAI, 2024a), o1-mini (OpenAI, 2024b), and Claude 3.5 Sonnet (Anthropic, 2024). GPT-4o and Claude 3.5 Sonnet are representative state-of-the-art language models; o1 and o1-mini are specifically tailored towards scientific reasoning during their training. We evaluate all four of these models as baselines to provide a comprehensive understanding of the performance of state-of-the-art LLMs.

## 5 Experimental Results

### 5.1 Main Results

Table 2 presents the performance of different methods, including their success rates and the qualities of their generated molecules. We include the full results in Table A3.

**LIDDIA *successfully generates novel, diverse, and high-quality molecules as potential drug candidates for* 73.3% *of targets (*TSR*), significantly outperforming existing methods.*** Pocket2Mol, the second-best method, achieves only a 23.3% success rate, while most proprietary LLMs fail entirely. Crucially, LIDDIA excels in simultaneously optimizing all five key pharmaceutical properties – QED, LRF, SAS, VNA, and NVT– on average, 85% of the generated molecules for each target are of high quality (HQ). In contrast, GPT-4o achieves only 35% in HQ, lagging nearly 50 percentage points behind LIDDIA, while all other methods perform even worse. In terms of the qualities of the generated molecules, LIDDIA produces molecules of comparable or superior quality to the limited outputs of other methods. These results highlight LIDDIA as a highly effective and reliable frame-

Table 2: Performance comparison between the baseline methods and LIDDIA. Full results is available in Table A3.

| | | Pocket2Mol | | DiffSMOL | | Claude | | GPT-4o | | o1-mini | | o1 | | LIDDIA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *%m/t* | *#m/t* | *%m/t* | *#m/t* | *%m/t* | *#m/t* | *%m/t* | *#m/t* | *%m/t* | *#m/t* | *%m/t* | *#m/t* | *%m/t* | *#m/t* |
| initial | `Generated` | - | 100.0 | - | 100.0 | - | 5.0 | - | 5.0 | - | 5.0 | - | 5.0 | - | 24.5 |
| | `Valid` | 100.0 | 100.0 | 99.9 | 99.9 | 98.7 | 4.9 | 97.3 | 4.9 | 91.3 | 4.6 | 95.3 | 4.8 | 100.0 | 24.5 |
| generated molecules | `QED` $\geq \overline{\texttt{QED}}_t$ | 53.4 | 53.4 | 60.0 | 60.0 | 96.7 | 4.8 | 88.2 | 4.4 | 90.1 | 4.5 | 88.3 | 4.4 | **97.2** | 21.8 |
| | `LRF` $\geq \overline{\texttt{LRF}}_t$ | **99.7** | 99.7 | 72.1 | 72.1 | 98.7 | 4.9 | 95.9 | 4.8 | 90.7 | 4.5 | 95.3 | 4.8 | 96.7 | 21.8 |
| | `SAS` $\leq \overline{\texttt{SAS}}_t$ | 77.4 | 77.4 | 7.5 | 7.5 | **92.7** | 4.6 | 90.7 | 4.5 | 81.4 | 4.1 | 92.6 | 4.6 | 88.3 | 17.4 |
| | `VNA` $\leq \overline{\texttt{VNA}}_t$ | 15.3 | 15.3 | 24.7 | 24.7 | 63.3 | 3.2 | 59.2 | 3.0 | 47.9 | 2.3 | 34.6 | 1.8 | **95.8** | 21.2 |
| | `NVT` $\geq 0.8$ | 87.6 | 87.6 | **98.2** | 98.2 | 46.9 | 2.4 | 68.3 | 3.4 | 64.1 | 3.2 | 55.9 | 2.8 | 97.8 | 22.4 |
| | `HQ` | 6.4 | 6.4 | 0.7 | 0.7 | 30.3 | 1.5 | 35.0 | 1.7 | 28.2 | 1.4 | 20.7 | 1.0 | **84.0** | 14.5 |
| | | *%t* | *#t* | *%t* | *#t* | *%t* | *#t* | *%t* | *#t* | *%t* | *#t* | *%t* | *#t* | *%t* | *#t* |
| among all targets | `DVS` $\geq 0.8$ | **100.0** | 30 | **100.0** | 30 | 30.0 | 9 | 90.0 | 27 | 67.7 | 20 | 70.0 | 21 | 97.7 | 29 |
| | $N \geq 5$&`DVS` | **100.0** | 30 | **100.0** | 30 | 27.7 | 8 | 77.7 | 23 | 43.3 | 13 | 57.7 | 17 | 90.0 | 27 |
| | $N \geq 5$&`HQ` | 27.7 | 8 | 3.3 | 1 | 23.3 | 7 | 10.0 | 3 | 0.0 | 0 | 3.3 | 1 | **73.3** | 22 |
| | `DVS` & `HQ` | 23.3 | 7 | 10.0 | 3 | 10.0 | 3 | 33.3 | 10 | 33.3 | 10 | 20.0 | 6 | **90** | 27 |
| | `TSR` | 23.3 | 7 | 0.0 | 0 | 6.7 | 2 | 6.7 | 2 | 0.0 | 0 | 0.0 | 0 | **73.3** | 22 |
| | Quality of Generated Molecules | | | | | | | | | | | | | | |
| | `NVT` ↑ | 0.87 | | **0.89** | | 0.77 | | 0.82 | | 0.79 | | 0.80 | | 0.86 | |
| | `QED` ↑ | 0.51 | | 0.55 | | **0.78** | | 0.74 | | 0.75 | | 0.77 | | 0.69 | |
| | `LRF` ↑ | **4.00** | | 3.43 | | **4.00** | | 3.99 | | 3.85 | | **4.00** | | 3.93 | |
| | `SAS` ↓ | 2.46 | | 6.15 | | 2.30 | | 2.16 | | **2.02** | | 2.03 | | 2.62 | |
| | `VNA` ↓ | -4.74 | | -4.23 | | -6.69 | | -6.56 | | -6.31 | | -5.97 | | **-7.17** | |
| | `DVS` ↑ | 0.88 | | **0.89** | | 0.76 | | 0.84 | | 0.79 | | 0.80 | | 0.82 | |

*%m/t*: average percentage of molecules per target; *#m/t*: average number of molecules per target; `Generated`: initially generated molecules; `Valid`: generated molecules that are also valid; overline$_t$: the average value of corresponding property in the known drugs for the target $t$. *%t*: average percentage of targets among all targets; *#t*: average number of targets; $N \geq 5$&`DVS`: at least 5 molecules are generated and the set is diverse; $N \geq 5$&`HQ`: at least 5 molecules are generated and they are of high quality; ↑/↓ indicates higher/lower values are better. **Bold** and underline indicates the best and second-best results, respectively.

work for accelerating drug discovery, consistently outperforming existing methods in both success rate and molecule qualities. We also compare LIDDIA with more recent state-of-the-art methods and observe similar findings as presented in Appendix C.1.

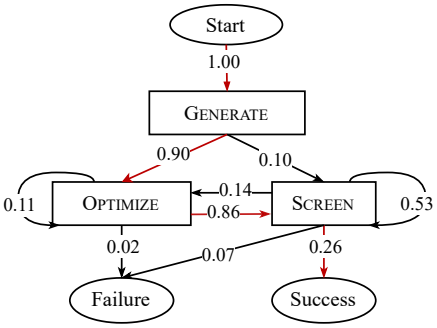



Figure 2: Action transitions in LIDDIA. The numbers represent the transition probabilities.

**LIDDIA *consistently generates high-quality molecules across all key pharmaceutical properties.*** Notably, it produces the most molecules (97.2%) for each target with `QED` higher than average, and the most molecules (95.8%) for each target with `VNA` higher than average, compared to other methods. Most of them (97.8%) are also novel, second only to DiffSMol. With respect to known drugs, a vast majority (88.3%) of LIDDIA's molecules exhibit high synthetic accessibility (`SAS`) and (96.7%) adhere to Lipinski's Rule of Five (`LRF`), comparable to general-purpose LLMs such as Claude, GPT-4o, and o1. On the most stringent metric, `VNA`, LIDDIA significantly outperforms other methods, with 95.8% of its generated molecules binding similarly to or better than existing drugs. In contrast, existing methods struggle to exceed 65% across these key properties. Overall, LIDDIA proves to be a robust and reliable approach, surpassing existing methods in generat-

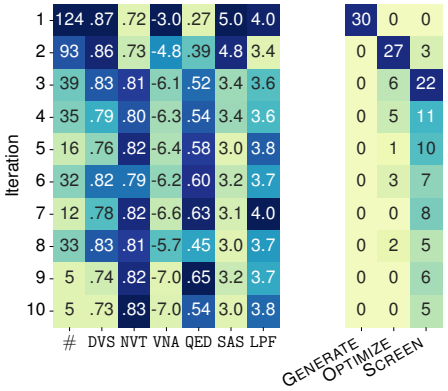



Figure 3: Molecule quality (left panel) and actions (right panel) over iterations by LIDDIA.

ing high-quality drug candidates. We also show that LIDDIA's generated molecules are better than or comparable to known drugs in terms of their toxicity properties. We provide further discussion in Appendix C.2.

***Existing methods face substantial challenges to achieve multiple good properties concurrently.*** For instance, general-purpose LLMs – Claude, GPT4o, o1-mini, and o1– exhibit a trade-off between novelty and binding. While they achieve impressive VNA, their generations tend to resemble known drugs closely (e.g., <70% novelty). In contrast, Pocket2mol and DiffSMol demonstrate the opposite: excel at NVT but struggle at VNA. It is possible that LLMs often anchor their generations based on prior knowledge (e.g., known ligands), thus narrowing their explorations. Meanwhile, Pocket2Mol and DiffSMol can generate new binding molecules but not better than existing drugs. Note that LIDDIA does not suffer from such compromise (e.g., both SAS and VNA >95%).

## 5.2 Agent Analysis

### 5.2.1 LIDDIA action patterns

Figure 2 presents the transition probabilities of actions that LIDDIA takes throughout the drug discovery process across all the targets, from start to finish.

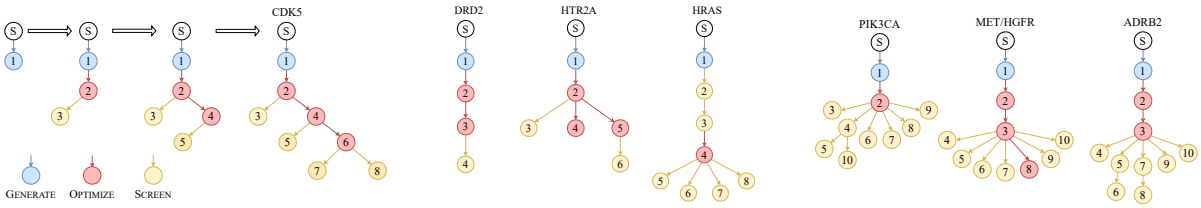
**LIDDIA** ***aligns with a typical drug discovery workflow, incorporating intelligent refinement at every stage.*** The most likely strategy of LIDDIA begins with the generation of target-binding molecules (GENERATE), followed by either optimization to enhance their properties (OPTIMIZE), or screening and selection over the generated molecules (SCREEN). Typically, optimization is necessary, which is followed by molecule screening over the optimized molecules. Iterative optimization is possible when no viable molecules exist. Similarly, iterative molecule screening is employed when plenty of viable molecules exist but are structurally similar. For instance, LIDDIA may cluster these molecules and subsequently identify the most promising molecules within each cluster. The most common workflow covers GENERATE, OPTIMIZE, and then SCREEN toward successful outcomes.

**SCREEN** ***serves as a quality-control mechanism to enable successful outcomes.*** Successful molecules are only possible after SCREEN completes screening and selection and identifies such molecules to output. As GENERATE and OPTIMIZE tend to yield more molecules than necessary, allowing abundant opportunities for LIDDIA to succeed, SCREEN prevents LIDDIA from producing low-quality drug candidates.

***Most of the generated molecules from*** **GENERATE** ***directly go through subsequent optimization by*** **OPTIMIZE.** This occurs in approximately 90% of the cases. Among all the generated molecules by GENERATE, it is typical that none of them satisfies all the property requirements, particularly those properties that are not integrated into the GENERATE tool designs. Any screening by SCREEN over such molecules will be futile and wasteful. Instead, LIDDIA intelligently executes OPTIMIZE, improving the likelihood of successful molecules out of SCREEN screening. Meanwhile, LIDDIA can still identify high-quality generated molecules and conducts screening directly over them. This clearly demonstrates the reasoning capability of LIDDIA as an effective tool for drug discovery.

**LIDDIA** ***leverages performance-driven insights to determine the most optimal action.*** Confronted with low-quality outputs from GENERATE, LIDDIA selectively pursues optimization to maximize results. Whenever additional molecules need to be considered (e.g., SCREEN does not identify good candidates), LIDDIA prioritizes optimizing promising molecules stored in MEMORY rather than generating entirely new ones. This represents a cost-effective, risk-averse strategy, balancing exploration and exploitation by refining known candidates with high potential rather than investing computational resources in *de novo* generation with suboptimal outcomes.

**LIDDIA** ***favors refining a few highly promising candidates as it continuously progresses.*** Figure 3 describes both the quality of the molecules produced and the typical actions LIDDIA takes at each step. Compared to the initial pool, the output molecules roughly achieve double the QED and VNA, thus emphasizing the importance of iterative optimization and effective screening strategies. However, diversity among these top candidates is often limited since molecules that satisfy multiple property requirements tend to converge on similar structures. This further highlights the complexity of drug discovery. Figure 3 (right) also highlights how LIDDIA tends to focus on refining the molecules via OPTIMIZE or SCREEN in later steps, mimicking a typical drug discovery workflow.




(a) Successful actions for CDK5 (b) Successful action trajectories (c) Failure action trajectories

Figure 4: LIDDIA actions trajectories across different targets.

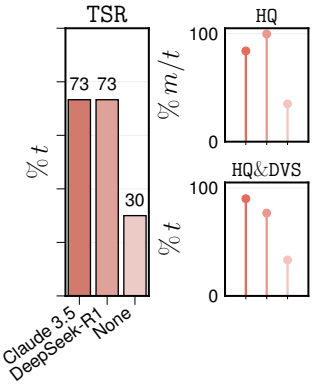
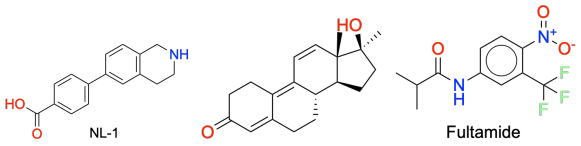
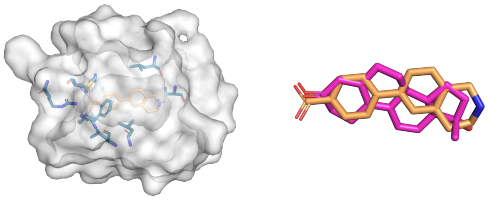



Figure 5: Ablation study on LIDDIA.




(a) (b) (c)

Figure 6: Case study on AR/NR3C4. (a) Promising molecule (NL-1) generated by LIDDIA. (b) Known ligand for AR/NR3C4. (c) An example of known approved drugs for AR/NR3C4.



(a) (b)

Figure 7: (a) Docking of NL-1 within the AR/NR3C4 pocket, with hydrogen bonds shown as solid blue lines. (b) NL-1 superpositioned with the known ligand for the AR/NR3C4 pocket. Orange denotes NL-1; Pink denotes the known ligand.

#### 5.2.2 Analysis on LIDDIA action trajectories

Figure 4 presents an example of LIDDIA action trajectories to identify promising drug candidates for CDK5 (Lau et al., 2002), a target for neurological conditions, such as Alzheimer's Dementia, along with examples of trajectories that lead to success (e.g., HTR2A, HRAS, DRD2) and fail (e.g., PIK3CA, MET, ADRB2) outcomes, respectively.

**LIDDIA *intelligently balances exploration and exploitation, critical to identify promising candidates***. In the case of CDK5, where it is highly challenging to identify a good drug candidate as demonstrated in the literature (Xie et al., 2022), LIDDIA is able to adaptively explore the chemical space via iteratively screening viable molecules and improving any property the molecules fail to meet. As shown in Figure 4 (a), LIDDIA starts with GENERATE (step 1), proceeds with OPTIMIZE (step 2), then applies SCREEN (step 3) but fails to find favorable candidates. In response, LIDDIA refines the failing property (step 4) and performs another screening (step 5). This process continues (steps 6, 7, 8) until the agent finally converges to a set of promising candidates.

This not an isolated case; LIDDIA consistently displays comparable intelligent decision-making behavior on other targets as observed in Figure 4 (b) and Figure 4 (c). Notably, in cases with successful outcomes, LIDDIA methodically refines several properties before screening for candidates (DRD2), or strategically determines which molecules to prioritize and what action to take (HTR2A and HRAS). For instance, in the HRAS case, LIDDIA uses several screenings (steps 2 and 3) to identify viable candidates, optimize them (step 4), and conduct further screenings (step 5 to 8) until it identifies favorable candidates. This highlights one key strength of LIDDIA– its capability to adapt to feedback (e.g., molecules quality) from its EVALUATOR, to explore (e.g., via refinement and generation), and to exploit (e.g., screening existing molecules) the chemical space. On cases where LIDDIA yields suboptimal results, such as PIK3CA, MET, and ADRB2, it still exhaustively performs various actions up to the action limits (i.e., 10 iterations). Additional analysis of PIK3CA, MET, and ADRB2 can be found in Appendix C.3.

### 5.3 Ablation Study

We perform additional experiments to test the effectiveness of LIDDIA. First, we replace Claude 3.5 Sonnet with DeepSeek-R1 in all components

requiring large language models to test LIDDIA's robustness to different backend LLMs.Additionally, we compare LIDDIA to a simple deterministic loop iterating between LIDDIA's components to analyze the importance of reasoning in LIDDIA. Note that this deterministic loop is similar to LIDDIA but without any LLM. Concretely, we run GENERATE once, followed by a loop of OPTIMIZE and SCREEN for $k$ number of times. We prioritize properties that fall below requirements when optimizing molecules, and only SCREEN for high-quality molecules. We set $k$ to 10, same as in Section 4.3. We show some results in Figure 5 and the full results in the Appendix (Table A3)

***Reasoning is critical for successful drug discovery***. LIDDIA with reasoning (both Claude 3.5 and DeepSeek-R1) achieves a much higher target success rate than without (more than 40% absolute difference), indicating its significance.

**LIDDIA *is robust to different backend LLMs.*** Comparing LIDDIA with Claude 3.5 and DeepSeek-R1, they both perform similarly (both with 73% `TSR`), emphasizing that our framework is robust to different backend LLMs.

Interestingly, molecules generated by LIDDIA with DeepSeek-R1 are almost always `HQ` compared to Claude (>99% vs 84%). However, only 77% satisfy the diversity requirements, in contrast to Claude 3.5 (90%).

### 5.4 Case Study on AR/NR3C4

We task LIDDIA with discovering new potential drug therapies targeting androgen receptor (AR/NR3C4), a hormone-driven transcription factor protein that plays a key role in both prostate and breast cancers (Tan et al., 2015; Giovannelli et al., 2018). LIDDIA identifies one molecule (named NL-1) with better `QED`, `VNA`, and `SAS` than the ligand and at least one approved drug (e.g., Fulmatide) for the respective targets. They are illustrated in Figure 6. NL-1 has several desirable traits, such as zwitterionic (with positive and negative charged atoms on the respective ends)—a trait typical in most biological molecules and drugs (Mobitz et al., 2024). The molecule also passes several computational filters, including PAINS (Baell and Holloway, 2010), BRENK (Brenk et al., 2008), NIH (Jadhav et al., 2010; Doveston et al., 2015), Lilly (Bruns and Watson, 2012), and Lipinski (Lipinski et al., 2001), further highlighting its attractiveness as a drug. In terms of binding, the molecule has -8.81 kcal/mol for `VNA`, emphasizing that it can bind well to the pocket. Notably, Figure 7a shows that the molecule is buried deep within the pocket, surrounded almost entirely by hydrophobic residues providing many van der Waals contacts. The molecule's carboxylic acid group also engages in hydrogen bonding at one end of the pocket, further stabilizing the complex and contributing to binding affinity. Encouragingly, further evaluation reveals that NL-1 has several established synthetic routes and demonstrates precedent for antagonizing stimulator of interferon genes (STING) (Gulen et al., 2023), thereby useful in the treatment of inflammatory diseases (Li et al., 2023). The desirable traits, combined with its synthetic accessibility and therapeutic precedents, position the molecule as a promising candidate for the androgen receptor.

Figure 7b shows a comparison of NL-1 to the known ligand. Both molecules feature a hydrophobic core with polar anchors on either end, but NL-1 is slightly less compact, with fewer fused rings than the known ligand. This reduces conformational rigidity, providing slightly increased flexibility to adapt to the binding pocket. Furthermore, NL-1 anchors itself with a carboxylic acid in place of the known ligand's ketone, enabling more hydrogen bonds.

We also present another case study on EGFR and extensively discuss the results in Appendix C.5.

## 6 Conclusions

We present LIDDIA, an agent for autonomous drug discovery. We comprehensively examine its capabilities, demonstrating its performance across many major therapeutical targets and revealing several key insights on its success. Furthermore, we investigate the generated molecules on the highly critical target EGFR and show their potential as drug candidates.

## 7 Acknowledgements

This project was made possible, in part, by support from the National Science Foundation grant nos. IIS-2133650 and IIS-2435819, and the National Library of Medicine grant no. 1R01LM014385. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agency. In addition, we would like to thank our colleagues, Ruoxi Gao, for generating some of the data and figures and Xinyi Ling for generating some of the figures in this work.

## 8 Limitations

LIDDIA is not without its limitations, and we plan to address them in future work. (1) We aim to show the utility of LLM in navigating drug discovery *in silico*, and thus, we solely focus on *in silico* evaluations. However, *in silico* is only a part of the entire drug discovery pipeline. To test LIDDIA's efficacy in the clinical world, we plan to add wet-lab validation in our follow-up research. (2) We focus our evaluation on a few key pharmaceutical properties without undermining others in our current work. We acknowledge that drug discovery requires much more than just a few key properties. The goal of this paper is not to replicate the entire drug discovery process but to demonstrate the strong potential of agents in facilitating drug discovery through generating and optimizing new drug candidates over a few essential properties. To this end, we intentionally design LIDDIA to be easily extendable to other metrics. We will continue developing the agent to cover more properties in our future research. (3) Our experiments with LIDDIA were built on a limited number of API calls due to budget constraints. Further testing on more API calls will be an interesting avenue for future research. (4) Finally, we benchmark LIDDIA on a small set of targets given the lack of well-established, large-scale, well-annotated benchmarks for our tasks. We prioritized a few highly clinically relevant therapeutic targets (such as cancer) with detailed information about their structures, ligands, and binding affinities. This serves as a trade-off between scope (lack of benchmarks) and substance (focusing on clinically relevant targets). This follows the example of related LLM agent works (Boiko et al., 2023; M. Bran et al., 2024), which have used similarly small benchmarks for initial demonstrations of agent capabilities on complex tasks. Future work will include an expansion of this dataset and additional benchmarking.

## Ethics Statement

LIDDIA is designed to generate small molecules meeting the parameters specified in a natural language prompt. However, we recognize that not all small molecules are safe, and such a tool could generate harmful molecules. As such, we have taken several steps to minimize the potential negative impact. LIDDIA only functions *in silico* and does not currently generate synthesis plans for any of its molecules. This prevents any dangerous molecules from being automatically produced without conscious human oversight or intervention. Additionally, our EVALUATOR implementation and example prompts focus LIDDIA on priority metrics in designing drugs to benefit humans, such as `QED` and `LRF`.

Despite these efforts, we cannot guarantee that LIDDIA will not generate harmful or incorrect content. We encourage users to practice discretion when using LIDDIA, and to follow all applicable safety guidelines and research best practices.

## References


Anthropic. 2024. Introducing Claude 3.5 Sonnet.

Jonathan B Baell and Georgina A Holloway. 2010. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry*, 53(7):2719–2740.

A. Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J. Bellis, Jon Chambers, Mark Davies, and 1 others. 2014. The ChEMBL bioactivity database: an update. *Nucleic Acids Research*, 42(Database issue):D1083–1090.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E Bourne. 2000. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. 2012. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98.

Benjamin E. Blass. 2021. Drug discovery and development: An overview of modern methods and principles. In Benjamin E. Blass, editor, *Basic Principles of Drug Discovery and Development (Second Edition)*, pages 1–41. Academic Press.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Ruth Brenk, Alessandro Schipani, Daniel James, Agata Krasowski, Ian Hugh Gilbert, Julie Frearson, and Paul Graham Wyatt. 2008. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3(3):435–444.

Robert F Bruns and Ian A Watson. 2012. Rules for identifying potentially reactive or promiscuous compounds. *Journal of medicinal chemistry*, 55(22):9763–9772.

Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. 2020. Retro*: Learning retrosynthetic planning with neural guided a* search. In *The 37th International Conference on Machine Learning (ICML 2020)*.

Ziqi Chen, Bo Peng, Tianhua Zhai, Daniel Adu-Ampratwum, and Xia Ning. 2025. Generating 3d small binding molecules using shape-conditioned diffusion models with guidance. *Nature Machine Intelligence*, pages 1–13.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, and 1 others. 2024. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*.

Debasis Das, Lingzhi Xie, and Jian Hong. 2024. Next-generation EGFR tyrosine kinase inhibitors to overcome C797S mutation in non-small cell lung cancer (2019–2024). *RSC Med. Chem.*, 15(10):3371–3394. Publisher: RSC.

Richard G Doveston, Paolo Tosatti, Mark Dow, Daniel J Foley, Ho Yin Li, Amanda J Campbell, David House, Ian Churcher, Stephen P Marsden, and Adam Nelson. 2015. A unified lead-oriented synthesis of over fifty molecular scaffolds. *Organic & biomolecular chemistry*, 13(3):859–865.

Maryne AJ Dubois, Rosemary A Croft, Yujie Ding, Chulho Choi, Dafydd R Owen, James A Bull, and James J Mousseau. 2021. Investigating 3, 3-diaryloxetanes as potential bioisosteres through matched molecular pair analysis. *RSC Medicinal Chemistry*, 12(12):2045–2052.

Peter Ertl and Ansgar Schuffenhauer. 2009. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8.

Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. 2020. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *J. Chem. Inf. Model.*, 60(9):4200–4215. Publisher: American Chemical Society.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with AI agents. *Cell*, 187(22):6125–6151. Publisher: Elsevier.

Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, and 1 others. 2011. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107.

Alireza Ghafarollahi and Markus J. Buehler. 2024. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 3(7):1389–1409. Publisher: Royal Society of Chemistry.

Pia Giovannelli, Marzia Di Donato, Giovanni Galasso, Erika Di Zazzo, Antonio Bilancio, and Antimo Migliaccio. 2018. The Androgen Receptor in Breast Cancer. *Frontiers in Endocrinology*, 9. Publisher: Frontiers.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 2023a. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. 2023b. Decompdiff: diffusion models with decomposed priors for structure-based drug design. In *Proceedings of the 40th International Conference on Machine Learning*, pages 11827–11846.

Muhammet F Gulen, Natasha Samson, Alexander Keller, Marius Schwabenland, Chong Liu, Selene Glück, Vivek V Thacker, Lucie Favre, Bastien Mangeat, Lona J Kroese, and 1 others. 2023. cgas–sting drives ageing-related inflammation and neurodegeneration. *Nature*, 620(7973):374–380.

Harold Hart. 1979. Simple enols. *Chemical Reviews*, 79(6):515–528.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, and 1 others. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37:19209–19253.

Yoshitaka Inoue, Tianci Song, and Tianfan Fu. 2024. DrugAgent: Explainable Drug Repurposing Agent with Large Language Model-based Reasoning. *arXiv preprint*. ArXiv:2408.13378.

Ajit Jadhav, Rafaela S Ferreira, Carleen Klumpp, Bryan T Mott, Christopher P Austin, James Inglese, Craig J Thomas, David J Maloney, Brian K Shoichet, and Anton Simeonov. 2010. Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *Journal of medicinal chemistry*, 53(1):37–51.

Jan H. Jensen. 2019. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chemical Science*, 10(12):3567–3572.

Lit-Fui Lau, Patricia A. Seymour, Mark A. Sanner, and Joel B. Schachter. 2002. Cdk5 as a Drug Target for the Treatment of Alzheimer's Disease. *Journal of Molecular Neuroscience*, 19(3):267–274.

Jing Li, Fengtao SONG, Sanjia XU, Wenqing Xu, and Zhiwei Wang. 2023. 3, 4-dihydroisoquinolin-1 (2h) -ones derivatives as sting antagonists and the use thereof.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1. *Advanced Drug Delivery Reviews*, 46(1):3–26.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and 1 others. 2023. Conversational Drug Editing Using Retrieval and Domain Feedback. In *ICLR 2024*.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535. Publisher: Nature Publishing Group.

Andrew D. McNaughton, Gautham Ramalaxmi, Agustin Kruel, Carter R. Knutson, Rohith A. Varikoti, and Neeraj Kumar. 2024. CACTUS: Chemistry Agent Connecting Tool-Usage to Science. *arXiv preprint*. ArXiv:2405.00972.

Henrik Mobitz, Birger Dittrich, Stephane Rodde, and Ross Strang. 2024. Nonclassical zwitterions as a design principle to reduce lipophilicity without impacting permeability. *Journal of medicinal chemistry*, 67(11):9485–9494.

H. L. Morgan. 1965. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113.

David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Jarrod Baker, Cinzia Malangone, and 1 others. 2023. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research*, 51(D1):D1353–D1359.

OpenAI. 2024a. Introducing OpenAI o1.

OpenAI. 2024b. OpenAI o1-mini.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, and 1 others. 2024. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. 2022. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17644–17655. PMLR. ISSN: 2640-3498.

RDKit. RDKit: Open-source cheminformatics.

Ansgar Schuffenhauer, Nadine Schneider, Samuel Hintermann, Douglas Auld, Jutta Blank, Simona Cotesta, Caroline Engeloch, Nikolas Fechner, Christoph Gaul, Jerome Giovannoni, and 1 others. 2020. Evolution of novartis' small molecule screening deck design. *Journal of medicinal chemistry*, 63(23):14425–14447.

Parthasarathy Seshacharyulu, Moorthy P Ponnusamy, Dhanya Haridas, Maneesh Jain, Apar K Ganti, and Surinder K Batra. 2012. Targeting the egfr signaling pathway in cancer therapy. *Expert Opinion on Therapeutic Targets*, 16(1):15–31. PMID: 22239438.

Ellen Swan, Kirsten Platts, and Anton Blencowe. 2019. An overview of the cycloaddition chemistry of fulvenes and emerging applications. *Beilstein Journal of Organic Chemistry*, 15(1):2113–2132.

Kyle Swanson, Parker Walther, Jeremy Leitz, Souhrid Mukherjee, Joseph C Wu, Rabindra V Shivnaraine, and 1 others. 2024. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7):btae416.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, and 1 others. 2023. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91. Publisher: Nature Publishing Group.

MH Eileen Tan, Jun Li, H. Eric Xu, Karsten Melcher, and Eu-leong Yong. 2015. Androgen receptor: structure, role in prostate cancer and drug discovery. *Acta Pharmacologica Sinica*, 36(1):3–23. Publisher: Nature Publishing Group.

Oleg Trott and Arthur J. Olson. 2010. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31(2):455–461.

Zhouling Xie, Shuzeng Hou, Xiaoxiao Yang, Yajun Duan, Jihong Han, Qin Wang, and Chenzhong Liao. 2022. Lessons Learned from Past Cyclin-Dependent Kinase Drug Discovery Efforts. *Journal of Medicinal Chemistry*, 65(9):6356–6389. Publisher: American Chemical Society.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR 2023*.

Jiajun Zhang and Xiao-Feng Wu. 2023. Palladium-catalyzed carbonylative synthesis of diaryl ketones from arenes and arylboronic acids through c (sp2)–h thianthrenation. *Organic Letters*, 25(12):2162–2166.

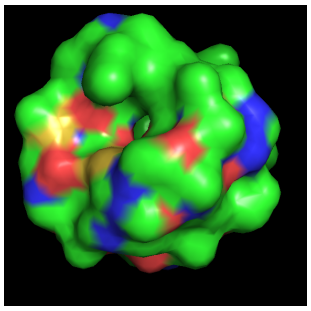
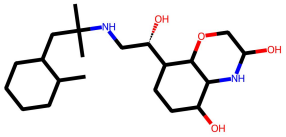
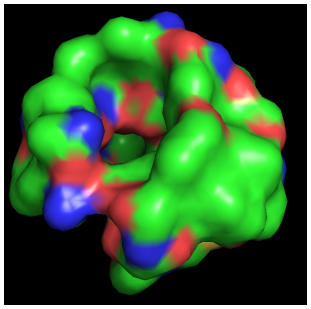
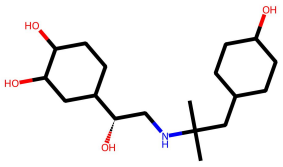
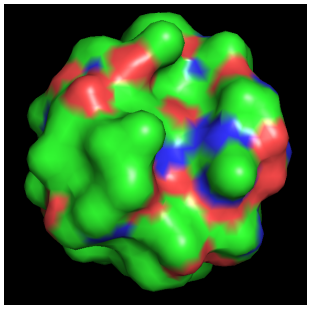
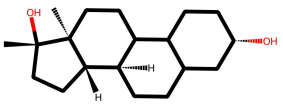
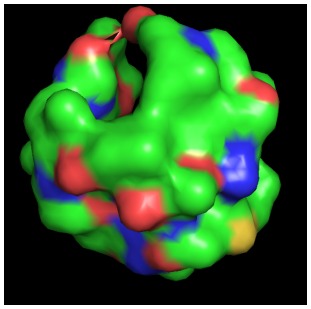
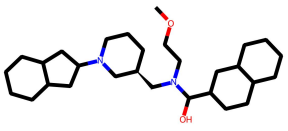
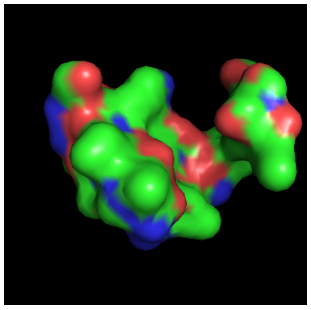
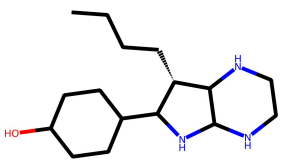
Dengwei Zhao, Shikui Tu, and Lei Xu. 2024. Efficient retrosynthetic planning with MCTS exploration enhanced A* search. *Commun Chem*, 7(1):1–12. Publisher: Nature Publishing Group.

Juexiao Zhou, Bin Zhang, Xiuying Chen, Haoyang Li, Xiaopeng Xu, Siyuan Chen, and 1 others. 2023. Automated Bioinformatics Analysis via AutoBA. *arXiv preprint*. ArXiv:2309.03242.

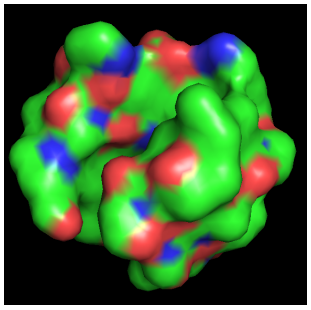
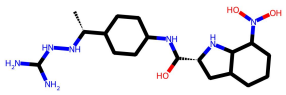
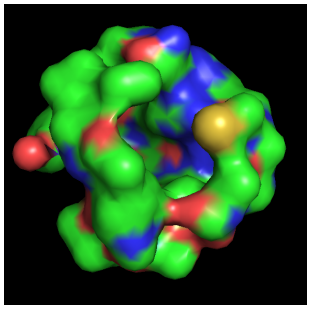
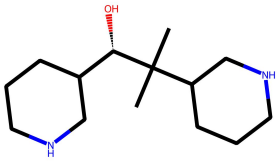
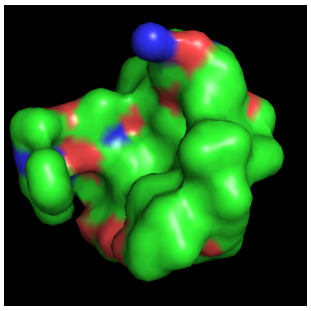
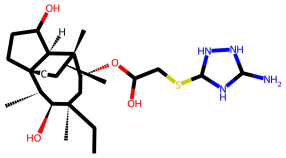
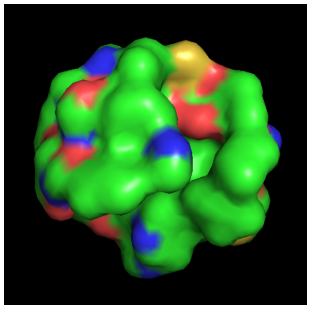
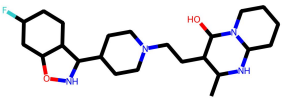
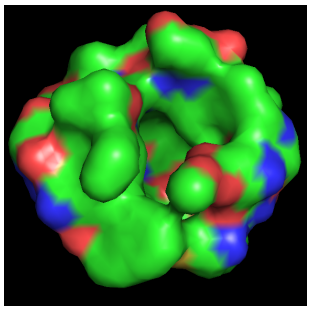
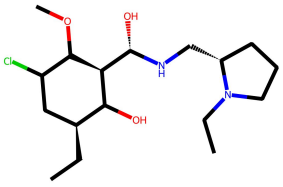
Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N. Zare, and Patrick Riley. 2019. Optimization of Molecules via Deep Reinforcement Learning. *Scientific Reports*, 9(1):10752. Publisher: Nature Publishing Group.

## A Dataset

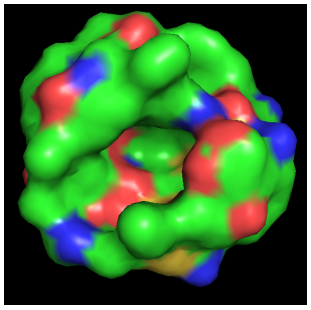
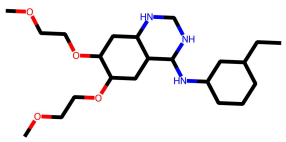
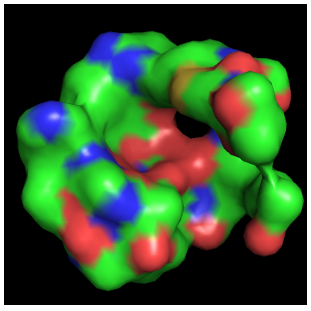
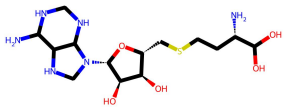
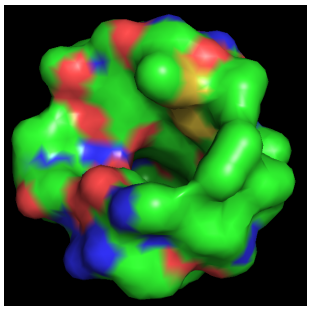
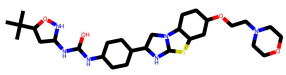
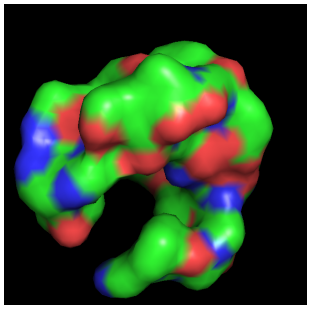
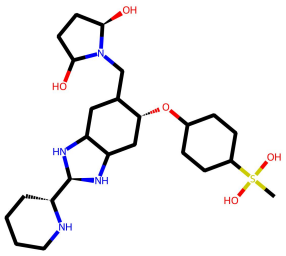
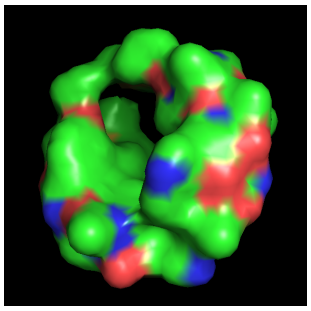
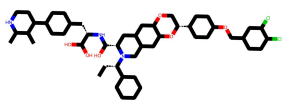
The following table details each protein target in the dataset. The disease categories column indicates whether the target is associated with autoimmune disease (A), cancer (CA), cardiovascular disease (CV), Diabetes (D), Infectious Disease (I), or Neurological Conditions (N).

| Gene Name | PDB ID | Pocket Structure | PDB Lig. | Ligand Structure | Disease Cat. |
|---|---|---|---|---|---|
| ADRB1 | 7BU7 |  | P0G |  | CV |
| ADRB2 | 4LDL |  | XQC |  | CV |
| AR (NR3C4) | 1E3G |  | R18 |  | CA |
| BCHE | 4TPK |  | 3F9 |  | N |
| CDK5 | 1UNG |  | ALH |  | N |

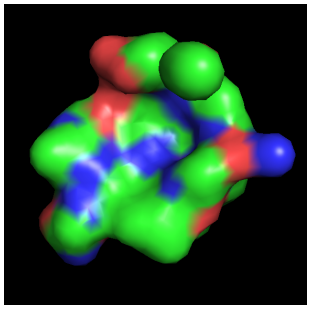
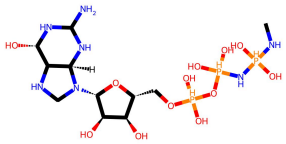
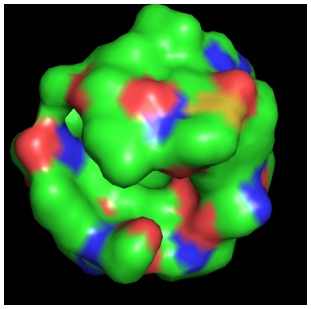
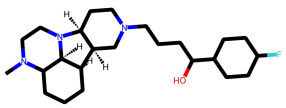
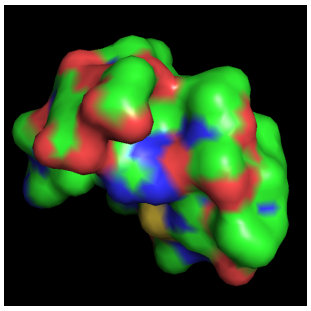
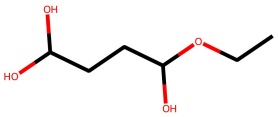
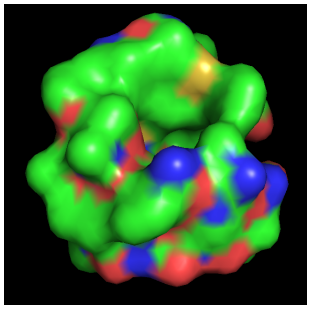
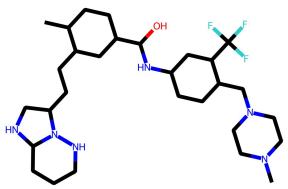
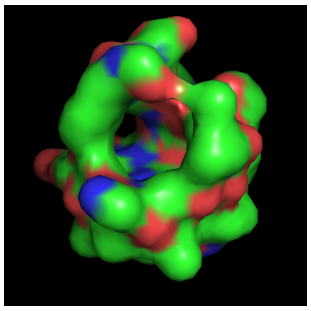
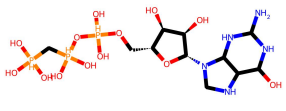
*continued*

| Gene Name | PDB ID | Pocket Structure | PDB Lig. | Ligand Structure | Disease Cat. |
|---|---|---|---|---|---|
| CHK2 | 2W7X |  | D1A |  | CA |
| CYP3A4 | 6MA6 |  | MYT |  | CA, I |
| CYP3A5 | 7SV2 |  | MWY |  | CA, I |
| DRD2 | 6CM4 |  | 8NU |  | N |
| DRD3 | 3PBL |  | ETQ |  | N |

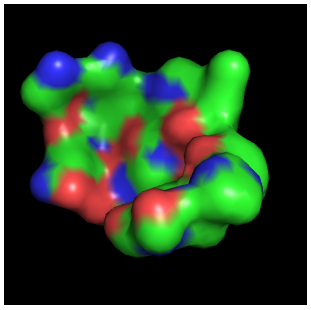
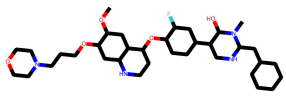
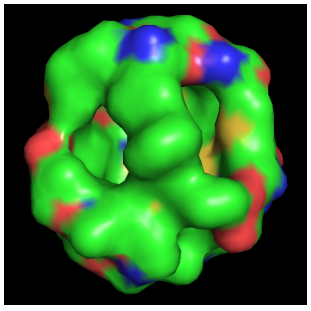
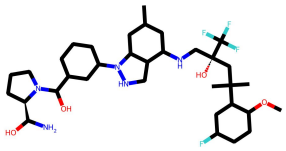
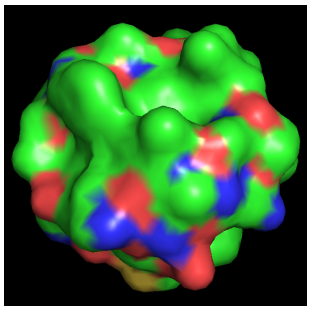
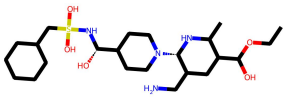
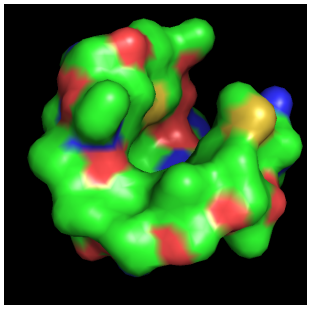
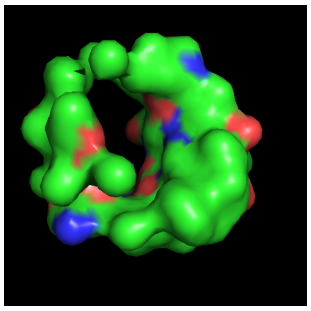
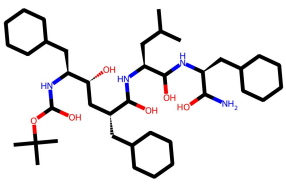
*continued*

| Gene Name | PDB ID | Pocket Structure | PDB Lig. | Ligand Structure | Disease Cat. |
|---|---|---|---|---|---|
| EGFR | 1M17 |  | AQ4 |  | CA |
| EZH2 | 7AT8 |  | SAH |  | CA |
| FLT3 | 4RT7 |  | P30 |  | CA |
| GCK (HK4) | 3H1V |  | GCK |  | D |
| GLP1R | 6ORV |  | N2V |  | D |

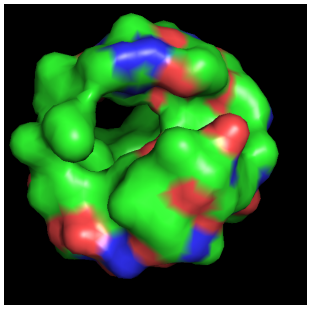
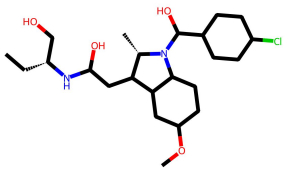
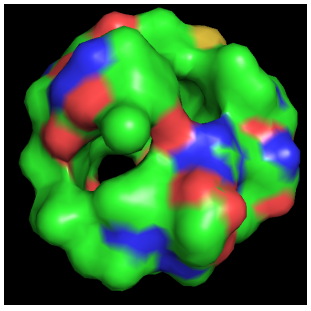
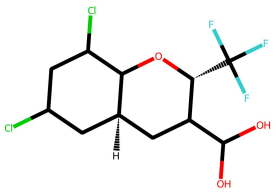
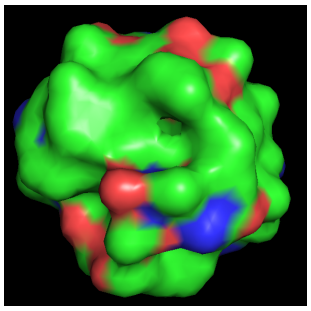
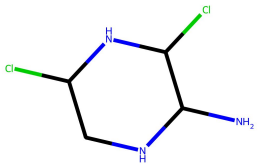
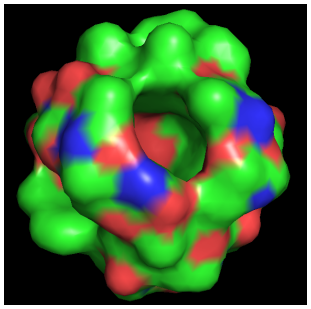
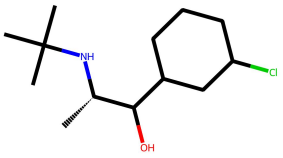
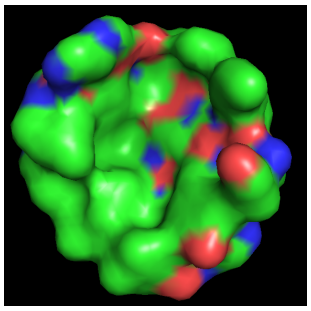
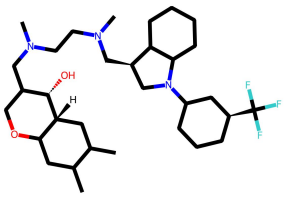
*continued*

| Gene Name | PDB ID | Pocket Structure | PDB Lig. | Ligand Structure | Disease Cat. |
|---|---|---|---|---|---|
| HRAS | 1RVD |  | DBG |  | CA |
| HTR2A | 7WC8 |  | 92S |  | N |
| KEAP1 | 7X4X |  | 9J3 |  | CA, N, A |
| KIT | 4U0I |  | 0LI |  | CA |
| KRAS | 4DSN |  | GCP |  | CA |

*continued*

| Gene Name | PDB ID | Pocket Structure | PDB Lig. | Ligand Structure | Disease Cat. |
|---|---|---|---|---|---|
| MET (HGFR) | 2RFN |  | AM7 |  | CA |
| NR3C1 (GR) | 3K23 |  | JZN |  | I, A |
| P2RY12 (P2Y12) | 4NTJ |  | AZJ |  | CV |
| PIK3CA | 4JPS |  | 1LT |  | CA |
| PSEN1 | 7C9I |  | FTO |  | N, CV |

*continued*

| Gene Name | PDB ID | Pocket Structure | PDB Lig. | Ligand Structure | Disease Cat. |
|---|---|---|---|---|---|
| PTGS1 (COX1) | 2OYE |  | IM8 |  | CV |
| PTGS2 (COX2) | 3LNO |  | 52B |  | CV |
| SHP2 (PTPN11) | 7GS9 |  | LV7 |  | CA, D |
| SLC6A2 | 8HFL |  | 1XR |  | N |
| TNF | 2AZ5 |  | 307 |  | CA, I, A |

### A.1 On the Necessity of Dataset Curation

Existing benchmark datasets for drug discovery (Guan et al., 2023b,a) contain non-human proteins. In fact, only around 40% of the test set are human proteins. Furthermore, the targets in this test set do not necessarily have existing drugs associated with them, giving us no reference to the multi-property requirements needed to identify `HQ` molecules. The lack of drugs associated with the targets is unsurprising,

since the benchmark was originally designed for binding pose and affinity prediction (Francoeur et al., 2020). Instead, we have decided to manually curate a set of 30 proteins, all of which (1) are human proteins, (2) are associated with major diseases, and (3) have known drugs targeting them.

## B Prompts

### B.1 REASONER Prompt

```
You have access to the following molecules and pockets:
{pocket_str}{mol_str}
You also have access to a set of actions:
{action_str}
Your job is to find molecules that satisfy these requirements:
{req_str}
Here is a history of actions you have taken and the results:
{history_str}
Here is the evaluation result from previous iteration:
{eval_str}

Let's think step by step and take your time before you answer the question. What is
the best action to take and what is the input of the action?

Remember that you currently have {resource_str} left to solve the task.
Remember that you can only use one action.

Your answer must follow this format:

Action: [name of action]
Input: [input of the action, should be the identifier like ['MOL001'] or
['POCKET001']]

If you plan to use "CODE" action, you need to include this additional format:

Desc: [explain what you want to do with the input of the action. Be as verbose and
descriptive as possible but at most three sentences. Always refer to the identifier
of the action input.]
```

### B.2 EVALUATOR Prompt

```
You have access to the following pool of molecules:
{mol_str}
Your job is to find molecules that satisfy these requirements:
{req_str}
Does this pool of molecules satisfy the requirements?
Remember that all molecules in the pool must satisfy the requirements.

Let's think step by step and answer with the following format:
Reason: (a compact and brief one-sentence reasoning)
Answer: (YES or NO)
```

### B.3 SCREEN Prompt

```
Your job is to make a Python function called _function.
The input is a Dict[str, pd.DataFrame] with the following columns:
["SMILES", "QED", "SAScore", "Lipinski", "Novelty", "Vina Score"].
The output must be a pandas DataFrame with the same columns as the input.
```

Table A2: Toxicity profiles of LIDDIA's generated molecules and known drugs. Lower indicates better profile.

| Toxicity profile | LIDDIA | Known Drugs |
|---|---|---|
| Mutagenicity | 0.32 | **0.27** |
| Carcinogenicity | **0.23** | 0.25 |
| Clinical Toxicity | **0.08** | 0.33 |
| DILI | 0.68 | **0.54** |
| hERG Blocking | **0.31** | 0.59 |
| Acute Toxicity | **2.45** | 2.78 |

```
The function should be able to do the following task: {input_desc}

Your output must follow the following format:

import pandas as pd

def _function(Dict[str, pd.DataFrame]) -> pd.DataFrame:
    #---IMPORT LIBRARIES HERE---#
    #---IMPORT LIBRARIES HERE---#

    #---CODE HERE---#
    #---CODE HERE---#

    output_df = ...
    return output_df

Make sure you import the necessary libraries.
```

# C Additional Results

## C.1 Additional Baselines

We compare LIDDIA with two more recent task-specific molecule generation methods, TargetDiff (Guan et al., 2023a) and DecompDiff (Guan et al., 2023b). We run similar experiments as in Table 2 and present the results in Table A3. Overall, we observe similar results to Section 5.1. First, both LIDDIA with DeepSeek-R1 and Claude significantly outperform both methods, highlighting the effectiveness of LIDDIA. Second, both methods also struggle to generate new binding molecules better than existing drugs, while LIDDIA does not.

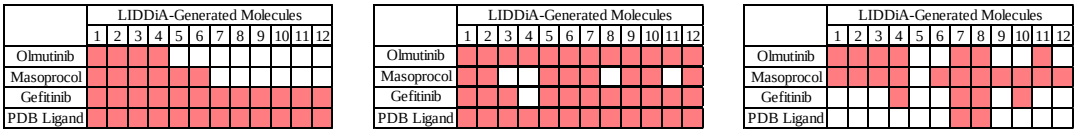
## C.2 Toxicity Predictions

We use ADMET-AI (Swanson et al., 2024) to predict the toxicity properties of LIDDIA's generated molecules. We use toxicity properties from Chen et.al (Chen et al., 2025) as reference and select some that are available in ADMET-AI. We then compare them to drugs in our dataset and present the results in Table A2.

Overall, we observe that our generated molecules are better than or comparable to known drugs in terms of their toxicity properties. Note that our agent is specifically designed to generate high-quality molecules, not "safe" molecules. Yet, Table A2 shows that their safety profiles are also promising. This presents interesting findings that (1) the current design of LIDDIA can generate both high-quality and "safe" molecules and (2) a significant opportunity for improvements to LIDDIA.

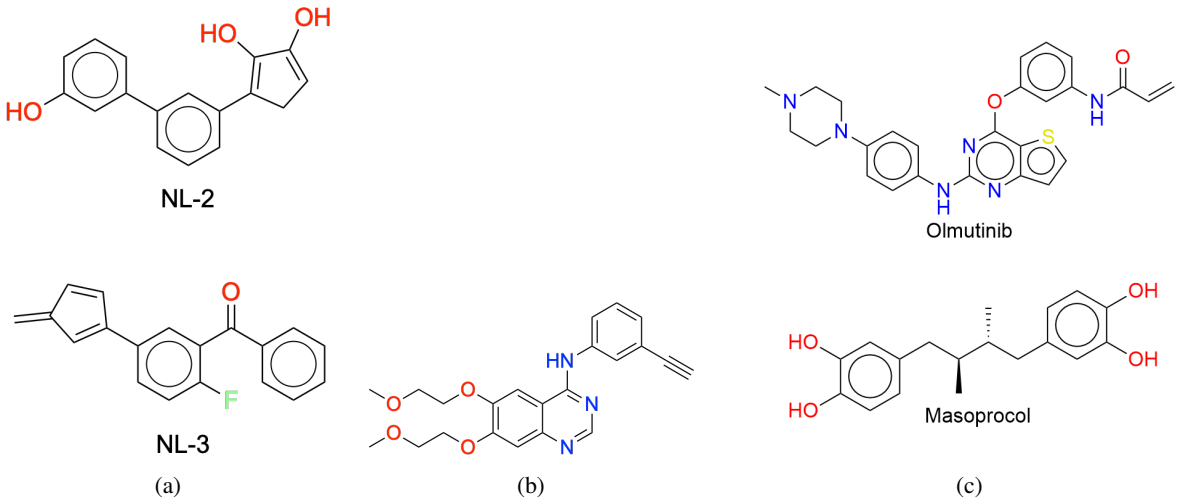
## C.3 Failure Analysis

We observe that LIDDIA yields suboptimal results on some targets (Section 5.2.2). First, we remark that on all targets, LIDDIA can generate at least one high-quality molecule, including on PIK3CA, MET, and ADRB2. However, they are suboptimal since: (1) the number of high-quality molecules is not sufficient




(a) VNA (b) QED (c) SAS

Figure A1: Case study for EGFR. Each subfigure compares molecules generated by LIDDIA to three drugs and one binding ligand of EGFR on VNA, SAS, and QED, respectively. Red squares indicate that the LIDDIA molecule outperforms the reference molecule on respective metrics.




(a) (b) (c)

Figure A2: Case study on EGFR. (a) LIDDIA's generated molecules (NL-2 and NL-3). (b) Known ligand for EGFR. (c) Examples of known approved drugs for EGFR. NL-2 has two enol groups and NL-3 has a fulvene, both of which are problematic as drug candidates.

(i.e., less than 5), or (2) the molecules are not diverse enough. We hypothesize that existing tools are struggling because of the structure of the pockets. For instance, the pocket may only allow a few specific scaffolds to bind, making it extremely difficult for existing tools to generate many and diverse high-quality molecules.

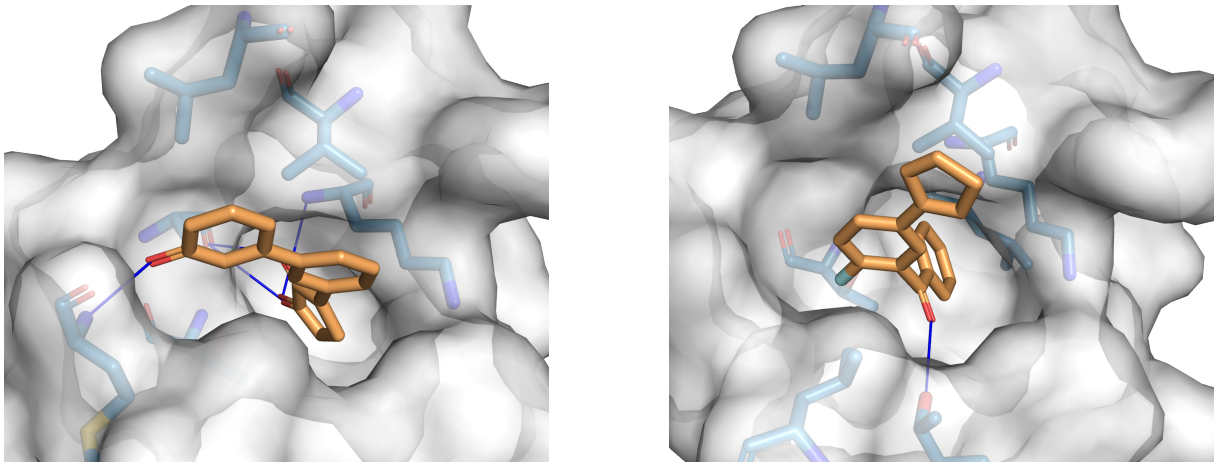
### C.4 Computational Costs

Overall, we observe that LIDDIA takes about 9K input tokens and 5K output tokens per target in our experiments. The estimated cost to generate high-quality molecules using LIDDIA is around $0.03 (USD) calculated using Claude pricing based on the number of used tokens. This highlights the potential of LIDDIA for low-cost autonomous drug discovery.

### C.5 Case Study on EGFR

We present another case study and task LIDDIA with discovering new potential drug therapies targeting the Epidermal growth factor receptor 1 (EGFR) protein. EGFR is a transmembrane glycoprotein that plays a pivotal role in many cancers, including breast cancer, esophageal cancer, and lung cancer (Seshacharyulu et al., 2012). Its role in cancer, as well as its accessibility on the cell membrane, has made it a prime therapeutic target (Bento et al., 2014; Gaulton et al., 2011). However, cancer cells mutate rapidly and can become resistant to drugs over time, leading to a need for novel drug therapies (Das et al., 2024). We compare the molecules generated for EGFR by LIDDIA with three approved drugs of EGFR – Olmutinib, Masoprocol, and Gefitinib, which exhibit the best VNA, QED and SAS among all EGFR's approved drugs, respectively. We also compare them with a known ligand for EGFR's binding pocket. Figure A1 presents the overall comparison results. Note that most existing methods cannot generate any novel high-quality molecules. This emphasizes the strength of LIDDIA, which can tackle even a challenging target.

**LIDDIA *effectively generates promising novel drug candidates on EGFR.*** Notably, these molecules surpass the native ligand in both VNA and QED, while displaying comparable overall profiles to approved



(a) (b)

Figure A3: (a) Docking of NL-2 to EGFR pocket, with hydrogen bonds shown as solid blue lines. (b) Docking of NL-3 to EGFR pocket, with hydrophobic van der Waals contacts shown using solid blue lines.

drugs. Moreover, some molecules (Figure A1 columns 1 and 2) are better than Olmutinib and Masoprocol on all metrics (i.e., `VNA`, `QED`, and `SAS`). We illustrate the two molecules, NL-2 and NL-3, in Figure A2. These molecules possess structural features that allow them to bind the pocket well. Notably, NL-2 has hydroxyl groups on both the five- and six-membered rings from the first molecule, which form strong hydrogen bonds with the protein target on opposite sides of the pocket (Figure A3a). NL-3 utilizes a different binding strategy, relying on hydrophobic packing and shape complementarity rather than polar interactions. As shown in Figure A3b, the fluorine substituent is positioned near the pocket entrance flanked by hydrophobic residues, which serve as favorable van der Waals contacts. Meanwhile, the diarylketone moiety is buried deep within the binding pocket, anchoring the ligand through planar stacking and hydrophobic interactions, despite the absence of direct hydrogen bonding. This finding aligns with previous literature (Zhang and Wu, 2023), which highlights the potency of diarylketone for antitumor drugs.

***Medicinal chemists are essential for a successful real-world deployment of* LIDDIA**. Despite the favorable binding and *in silico* properties, closer examination reveals some concerning structural features in these molecules. NL-2 contains two enol groups (the -OH near the double bond) —substructures with tautomeric instability and are highly unattractive for drugs (Hart, 1979). NL-3 contains fulvene, known to be chemically reactive, thermally unstable, sensitive to oxygen, and photosensitive (Swan et al., 2019). The diarylketone moiety, despite its favorable binding and potency in antitumor drugs (Zhang and Wu, 2023), is known to be phototoxic (Dubois et al., 2021). Such conflicts (e.g., favorable binding but phototoxic) are typical in drug discovery, highlighting its significant challenges and the necessity for more comprehensive evaluation for a successful practical deployment of LIDDIA.

Furthermore, no standalone *in silico* evaluation tools (e.g., computational filters from RDKit (RDKit) and Medchem (Schuffenhauer et al., 2020)) can detect all the issues presented in these molecules. Several filters (e.g., PAINS (Baell and Holloway, 2010), BRENK (Brenk et al., 2008), NIH (Jadhav et al., 2010; Doveston et al., 2015)) cannot capture the problematic features in NL-2, highlighting the limitation of existing tools. Lilly rules (Bruns and Watson, 2012) are able to identify the enol groups, but do not raise any alerts for NL-3. These findings underscore that human expertise remains irreplaceable in drug discovery—a domain where nuanced understanding and reliable assessments are critical for mitigating risks. They also highlight three priorities for the future work of LIDDIA: (1) human-in-the-loop validation, (2) development and integration of more sophisticated *in silico* tools, and (3) wet-lab validation of generated molecules.

As discussed, the lack of nuanced understanding by existing *in silico* tools contribute to the problematic features existed in LIDDIA's pool of generated molecules. A more reliable option is the inclusion of human experts in the loop for validation for a more nuanced and comprehensive evaluation of the

molecules. Meanwhile, existing *in silico* tools, particularly in evaluation, have rooms for improvement. The integration of more and better state-of-the-art tools can certainly benefit LIDDIA in generating more and better high-quality molecules. Ultimately, *in vitro* and *in vivo* in a laboratary will be necessary to analyze how LIDDIA's performance translates to real-world impacts. Thus, though encouraging, LIDDIA calls for more comprehensive and systematic investigation for a successful practical deployment.

Table A3: Performance comparison between task-specific molecule generation methods (Pocket2Mol, DiffSMol, TargetDiff, DecompDiff), general-purpose LLMs (Claude, GPT-4o, o1-mini and o1), and LIDDIA variants.

| | | Pocket2Mol | | DiffSMOL | | TargetDiff* | | DecompDiff | | Claude | | GPT-4o | | o1-mini | | o1 | | LIDDIA (None)† | | LIDDIA (DeepSeek) | | LIDDIA (Claude) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ | $\%m/t$ | $\#m/t$ |
| initial | `Generated` | - | 100.0 | - | 100.0 | - | 100.0 | - | 100.0 | - | 5.0 | - | 5.0 | - | 5.0 | - | 5.0 | - | 100.0 | - | 25.1 | - | 24.5 |
| | `Valid` | 100.0 | 100.0 | 99.9 | 99.9 | 87.5 | 87.5 | 78.5 | 78.5 | 98.7 | 4.9 | 97.3 | 4.9 | 91.3 | 4.6 | 95.3 | 4.8 | 100.0 | 100.0 | 100.0 | 25.1 | 100.0 | 24.5 |
| generated molecules | `QED` $\geq \overline{\texttt{QED}}_t$ | 53.4 | 53.4 | 60.0 | 60.0 | 47.3 | 47.3 | 39.1 | 39.1 | 96.7 | 4.8 | 88.2 | 4.4 | 90.1 | 4.5 | 88.3 | 4.4 | 81.6 | 81.6 | **100.0** | 25.1 | 97.2 | 21.8 |
| | `LRF` $\geq \overline{\texttt{LRF}}_t$ | **99.7** | 99.7 | 72.1 | 72.1 | 80.2 | 80.2 | 56.3 | 56.3 | 98.7 | 4.9 | 95.9 | 4.8 | 90.7 | 4.5 | 95.3 | 4.8 | 87.6 | 87.6 | **100.0** | 25.1 | 96.7 | 21.8 |
| | `SAS` $\leq \overline{\texttt{SAS}}_t$ | 77.4 | 77.4 | 7.5 | 7.5 | 22.3 | 22.3 | 13.4 | 13.4 | **92.7** | 4.6 | 90.7 | 4.5 | 81.4 | 4.1 | 92.6 | 4.6 | 69.6 | 69.6 | **100.0** | 25.1 | 88.3 | 17.4 |
| | `VNA` $\leq \overline{\texttt{VNA}}_t$ | 15.3 | 15.3 | 24.7 | 24.7 | 19.5 | 19.5 | 26.4 | 26.4 | 63.3 | 3.2 | 59.2 | 3.0 | 47.9 | 2.3 | 34.6 | 1.8 | 80.7 | 80.7 | **99.7** | 24.8 | 95.8 | 21.2 |
| | `NVT` $\geq 0.8$ | 87.6 | 87.6 | **98.2** | 98.2 | 82.2 | 82.2 | 70.1 | 70.1 | 46.9 | 2.4 | 68.3 | 3.4 | 64.1 | 3.2 | 55.9 | 2.8 | 83.9 | 83.9 | **100.0** | 25.1 | 97.8 | 22.4 |
| | `HQ` | 6.4 | 6.4 | 0.7 | 0.7 | 1.9 | 1.9 | 1.2 | 1.2 | 30.3 | 1.5 | 35.0 | 1.7 | 28.2 | 1.4 | 20.7 | 1.0 | 35.0 | 35.0 | **99.7** | 24.8 | 84.0 | 14.5 |
| | | $\%t$ | $\#t$ | $\%t$ | $\#t$ | $\%t$ | $\#t$ | $\%t$ | $\#t$ | $\%t$ | $\#t$ | $\%t$ | $\#t$ | $\%t$ | $\#t$ | $\%t$ | $\#t$ | $\%t$ | $\#t$ | $\%t$ | $\#t$ | $\%t$ | $\#t$ |
| among all targets | `DVS` $\geq 0.8$ | **100.0** | 30 | **100.0** | 30 | **100.0** | 29 | 97.7 | 29 | 30.0 | 9 | 90.0 | 27 | 67.7 | 20 | 70.0 | 21 | 70.0 | 21 | 76.7 | 23 | 97.7 | 29 |
| | $N \geq 5$ & `DVS` | **100.0** | 30 | **100.0** | 30 | **100.0** | 29 | 97.7 | 29 | 27.7 | 8 | 77.7 | 23 | 43.3 | 13 | 57.7 | 17 | 70.0 | 21 | 73.3 | 22 | 90.0 | 27 |
| | $N \geq 5$ & `HQ` | 27.7 | 8 | 3.3 | 1 | 13.8 | 4 | 13.3 | 4 | 23.3 | 7 | 10.0 | 3 | 0.0 | 0 | 3.3 | 1 | 93.3 | 28 | **97.7** | 29 | 73.3 | 22 |
| | `DVS` & `HQ` | 23.3 | 7 | 10.0 | 3 | 24.1 | 7 | 17.7 | 5 | 10.0 | 3 | 33.3 | 10 | 33.3 | 10 | 20.0 | 6 | 33.3 | 10 | 76.7 | 23 | **90** | 27 |
| | `TSR` | 23.3 | 7 | 0.0 | 0 | 13.8 | 4 | 13.3 | 4 | 6.7 | 2 | 6.7 | 2 | 0.0 | 0 | 0.0 | 0 | 30.0 | 9 | **73.3** | 22 | **73.3** | 22 |
| | Quality of Generated Molecules | | | | | | | | | | | | | | | | | | | | | | |
| | `NVT` ↑ | 0.87 | | **0.89** | | 0.88 | | 0.86 | | 0.77 | | 0.82 | | 0.79 | | 0.80 | | 0.85 | | 0.86 | | 0.86 | |
| | `QED` ↑ | 0.51 | | 0.55 | | 0.51 | | 0.47 | | **0.78** | | 0.74 | | 0.75 | | 0.77 | | 0.69 | | 0.70 | | 0.69 | |
| | `LRF` ↑ | **4.00** | | 3.43 | | 3.76 | | 3.20 | | **4.00** | | 3.99 | | 3.85 | | **4.00** | | 3.82 | | 3.97 | | 3.93 | |
| | `SAS` ↓ | 2.46 | | 6.15 | | 4.23 | | 4.73 | | 2.30 | | 2.16 | | **2.02** | | 2.03 | | 2.79 | | 2.49 | | 2.62 | |
| | `VNA` ↓ | -4.74 | | -4.23 | | -4.94 | | -4.88 | | -6.69 | | -6.56 | | -6.31 | | -5.97 | | -7.43 | | -7.24 | | **-7.17** | |
| | `DVS` ↑ | 0.88 | | **0.89** | | **0.91** | | 0.88 | | 0.76 | | 0.84 | | 0.79 | | 0.80 | | 0.81 | | 0.81 | | 0.82 | |

$\%m/t$: average percentage of molecules per target; $\#m/t$: average number of molecules per target; `Generated`: initially generated molecules; `Valid`: generated molecules that are also valid; $\overline{\text{overline}}_t$: the average value of corresponding property in the known drugs for the target $t$. $\%t$: average percentage of targets among all targets; $\#t$: average number of targets; $N \geq 5$&`DVS`: at least 5 molecules are generated and the set is diverse; $N \geq 5$&`HQ`: at least 5 molecules are generated and they are of high quality; ↑/↓ indicates higher/lower values are better. **Bold** and underline indicates the best and second-best results, respectively. *We run TargetDiff only on 29 pockets, as we cannot generate any molecules in one of the pockets (EZH2) due to some bugs in the code. † This method represents the "simple deterministic loop" described in the ablation study or LIDDIA without reasoning.