Towards LLM-Driven Multi-Agent Pipeline for Drug Discovery: Neurodegenerative Diseases Case Study

Gleb V. Solovev[1], Alina B. Zhidkovskaya[1], Anastasia Orlova[1], Anastasia Vepreva[1], Ilya Tonkii[1], Rodion Golovinskii[1], Nina Gubina[1], Denis Chistiakov[1], Timur A. Aliev[1], Ivan Poddiakov[2], Galina Zubkova[2], Ekaterina V. Skorb[1], Vladimir Vinogradov[1], Nikolay Nikitin[1], Andrei Dmitrenko[1], Anna Kalyuzhnaya[1], Andrey V. Savchenko[2,3]

[1]ITMO University, Saint Petersburg, Russia
[2]Sber AI Lab, Moscow, Russia
[3]HSE University, Moscow, Russia

Abstract

Recent studies demonstrate that Large Language Models (LLMs) can accelerate scientific progress in chemistry and drug development. However, existing approaches have not achieved successful automation of the complete drug discovery pipeline, primarily due to the absence of comprehensive datasets and the limitations of single-model solutions. This paper introduces multi-agent approach that combines LLMs with specialized generative models and validation tools to automate the end-to-end drug discovery process. The key innovation lies in addressing the complex transition from natural language problem formulation to building a complete computational pipeline for real pharmaceutical research tasks. Experimental results demonstrate that our multi-agent solution achieves 92% accuracy in end-to-end drug search complex tasks, significantly outperforming single-agent implementations. We validated the system's effectiveness on an original newly farmed dataset with tasks and full solutions for three pharmaceutical cases targeting neurodegenerative diseases (Alzheimer's, multiple sclerosis, and Parkinson's). The main contributions include demonstrating the advantages of a multi-agent LLM-powered approach for automating pharmaceutical drug design and validating its success on real-world drug discovery challenges.

Introduction

Drug discovery remains one of the pharmaceutical industry's most time-consuming and expensive processes (Huang et al. 2024). The automation of this process represents a critical challenge for modern medicine. While various computational molecular design approaches exist (Zeng et al. 2022), there remains a significant gap between human-interpretable drug requests for certain purposes with specific properties and the automated generation of highly valuable candidate molecules.

This challenge is particularly acute for neurodegenerative diseases like Alzheimer's, Parkinson's, and multiple sclerosis, which affect millions globally (Pushkaran and Arabi 2024). Neurological disorders are among the most challenging to treat because of the unique biological and chemical hurdles associated with the brain. The blood-brain barrier (BBB) restricts the passage of many therapeutic agents, necessitating the design of molecules with specific physicochemical properties to ensure effective delivery (Wu et al. 2023). Additionally, the multifactorial nature of these diseases often requires therapies that can target multiple pathways simultaneously, adding to the complexity of the drug discovery process (Jha et al. 2022).

Large Language Models (LLMs) have emerged as powerful tools for bridging this gap, offering impressive capabilities in understanding natural language descriptions of desired molecular properties and translating them into actionable parameters (Guan and Wang 2024). Several promising specialized LLM-based tools exist, e.g. DrugLLM (Liu et al. 2024) for zero-shot molecular generation and ChemLLM (Zhang et al. 2024a) for chemistry-related Q&A. However, the direct application of pre-trained and even fine-tuned LLMs to chemistry tasks is limited by their inability to perform all required specialized operations (such as molecular generation, validation, filtering, and properties prediction) and execute them in proper order.

To overcome these limitations, we propose a multi-agent approach that combines LLMs' natural language understanding capabilities with highly specialized generative models and validation tools for drug discovery tasks. Our system employs specialized agents with different cognitive functions: Planner Agent, Tool-calling Agent, Validator Agent and Summarizer Agent. These agents transform unstructured textual descriptions into valid molecular structures with desired properties. This approach enables end-to-end automation of the drug discovery process, from initial property specification to final molecule generation and validation.

The key challenge we address is the complex transition from problem formulation to building a complete computational pipeline for drug discovery for real pharmaceutical research tasks instead of academic examples. That is why we formulate the main research hypothesis as follows:

A multi-agent LLM approach can automate the full drug discovery pipeline from natural language task formulation to valid molecular candidates for real pharmaceutical research tasks significantly better than single-agent approach.

Our experimental results support these hypotheses. The proposed multi-agent solution achieves 92% accuracy[1] in

[1]This accuracy indicates how correctly the agents perform the user's query.

end-to-end drug search tasks, significantly outperforming a single-agent implementation (e.g., (M. Bran et al. 2024)) (71%). In complex queries of 4-5 tasks, the accuracy of the single-agent pipeline is zero - it cannot handle such a large amount of tasks , while the multi-agent pipeline shows 92% as well. We experimentally validated the effectiveness of the pipeline by testing it on three pharmaceutical cases searching for drugs to treat neurodegenerative diseases (Alzheimer's disease, multiple sclerosis, Parkinson's disease).

The main *contributions* of the paper are: (1) demonstration of advantage of multi-agent LLM-powered approach for automation pharmaceutical drug design research tasks; (2) demonstration of successfulness of suggested approach for real-world drug design challenges.

Related Work

ML models for drug design

Drug design is a rapidly growing field combining chemistry and machine learning. Traditionally, discovering new molecules or selecting chemical structures to solve a particular problem relies on existing experimental data and subjective research experience, which limits the number and variety of possible compounds that can be considered. Generative models allow efficient exploration of the molecular space, which has already fueled the explosive growth of molecular generative design. Recurrent neural networks (Grisoni et al. 2020; Li et al. 2020; Suresh et al. 2022; Dollar et al. 2021), variational autoencoders (Gómez-Bombarelli et al. 2018; Lee and Min 2022; Ochiai et al. 2023; Bhadwal, Kumar, and Kumar 2023), generative-adversarial networks (Guimaraes et al. 2017; Prykhodko et al. 2019; Pang et al. 2023; Macedo, Ribeiro Vaz, and Taveira Gomes 2024), evolutionary algorithms (Yoshikawa et al. 2018; Leguy et al. 2020; Kerstjens and De Winter 2022; Jensen 2019; Tripp and Hernández-Lobato 2023), and hybrid models using reinforcement learning methods (Putin et al. 2018; Thomas et al. 2022; Zhavoronkov et al. 2019) have been successfully applied to solve various problems in chemistry.

Another advanced model for sequence generation is Transformer (Vaswani et al. 2017), which is based on the attention mechanism. For molecule generation task, this model has successfully shown high performance in several studies (Ang, Rakovski, and Atamian 2024; Mao et al. 2023; Haroon, Hafsath, and Jereesh 2023). Researchers attribute this architecture's high performance to its ability to handle long sequences, which is applicable for chemical structures as they are usually treated as a sequence of atoms and bonds.

Despite various powerful methods for molecule generation, developing drugs to treat brain diseases still poses many challenges (Khachaturian et al. 2023). To date, researchers believe that machine learning and deep learning tools can effectively solve such complex problems(Vicidomini et al. 2024; Doherty et al. 2023).

Chemical LLM pipelines

Agent-based pipelines have gained widespread adoption in chemistry and pharmacology tasks (M. Bran et al. 2024; Zhang et al. 2024b; McNaughton et al. 2024; Li et al. 2024b; Jablonka et al. 2023) since late 2023. These systems automate experiments, significantly reducing time and financial costs and enabling professionals to achieve their objectives more efficiently.

Chemical agent systems can address a broad range of tasks, including prediction and modeling of chemical reactions; calculation of physical and chemical properties of substances; generation of molecules with specific characteristics; optimization of molecular structures to enhance their efficacy or reduce toxicity; analysis and prediction of pharmacokinetic and pharmacodynamic properties of compounds; assessment of the likelihood of successful synthesis for new compounds; automation of testing and selection processes for target molecules.

For example, the ChemCrow (M. Bran et al. 2024) is based on autonomous planning and execution of chemical synthesis using a robotics platform. This solution supports 15 applied tasks in chemistry, including the planning and synthesis of a repellent (DEET), the search for and synthesis of a thiourea organocatalyst for the Diels-Alder reaction, and the synthesis of paracetamol, aspirin, Safinamide, and Atorvastatin. Another example is chemical agent CACTUS (McNaughton et al. 2024) which is able to solve tasks such as molecular property prediction, similarity searching, and drug-likeness assessment but it supports only one-step tasks that seems to be not enough for many real-world research tasks.

Another example is the PhysicsAssistant platform (Latif, Parasuraman, and Zhai 2024). This platform employs LLMs to facilitate interactive learning in physics, helping students conduct experiments and analyze results. Similarly, CancerGPT, a model for predicting drug pair synergy using few-shot learning, was introduced to accelerate the development of new therapies (Li et al. 2024b).

Interest in using LLMs in chemistry is growing as they demonstrate potential in predictive analytics, molecular modeling, and developing new compounds. For instance, Chemdfm, a conversational platform powered by LLMs, was proposed for working with chemical data (Zhao et al. 2024). Research by M. Bran A. et al. (M. Bran et al. 2024) showed that integrating LLMs with chemical tools improves molecular property predictions. Finally, Ye G. (Ye 2024) proposed a novel approach for de novo drug design using LLMs, enabling the automation of new chemical compound generation.

These studies highlight that LLMs can accelerate scientific progress in chemistry and related fields, unlocking new material and drug development opportunities. Nevertheless, none of them demonstrate successful automation of full drug discovery pipeline. Possible reasons are (1) absence of easy accessible dataset with full drug discovery pipeline for training and validation of new models and approaches, (2) weakness of existed separated models and agents for solution the full task with high quality level. Based on these conclusions we propose our vision of stronger approach and high valuable newly farmed dataset with state-of-the-art pharmaceutical research task and their solutions.

Proposed Multi-Agent Approach

Drug discovery is a complex task that involves multiple processing tasks: understanding user requirements, generating valid molecular structures, predicting their properties, and evaluating results. Single-LLM approaches (e.g. DrugLLM (Liu et al. 2024) as well, as other chemistry-specific models (Grisoni 2023)) and single LLM-agents ((M. Bran et al. 2024), (McNaughton et al. 2024)) struggle to handle this complexity (Garg 2024; Allenspach, Hiss, and Schneider 2024).

Based on our experimental findings and literature analysis, we propose a multi-agent architecture (Figure 1) that addresses the inherent complexity of drug discovery pipeline automation. Our empirical study also shows that single-agent approaches struggle with complex queries requiring diverse tool interactions and domain-specific processing.

The system's architecture separates key cognitive functions into specialized agents, each optimized for specific tasks in the drug discovery workflow. The system utilizes four agents, each performing a unique task to achieve high operational accuracy (pipeline planning, governing of the strategy for tools usage, molecules validation, pipeline results summarization). This separation is grounded in modular software design principles and reduction of actions' space for each agent.

We observed that such a modular approach provides at least two advantages: (1) it allows each agent to focus on a narrower task domain, improving accuracy and reducing error propagation, and (2) it enables easier scalability in a case of multi-step and multi-turn pipelines without dramatic loss of quality.

Agents

The pipeline receives a textual query from the user, passed to the Planner Agent. In a case of complex tasks, this agent decomposes an initial query into more simple subtasks and maps each subtask to a specific query type for subsequent processing by a Tool-calling Agent. If the query is ambiguous, the Planner Agent is invoked to refine it and enrich its context with additional data the user provides.

The implementation includes a Tool-calling Agent increases a quality of calling the required models and functions.

The Tool-calling Agent determines required parameters and executes tool calls in JSON format. For molecule generation tasks, it interfaces with our generative models to produce candidate structures matching specified properties.

Validator Agent evaluates tool outputs against quality criteria, triggering reruns via the Planner if results are unsatisfactory. This ensures generated molecules meet all specified constraints.

Summarizer Agent compiles verified outputs into a cohesive response, presenting generated molecules with their calculated properties in a structured format.

This architecture enables end-to-end automation of the drug discovery pipeline while maintaining result quality through validation loops. Such system can handle both simple single-task queries and complex multi-step requests requiring coordination between different tools.

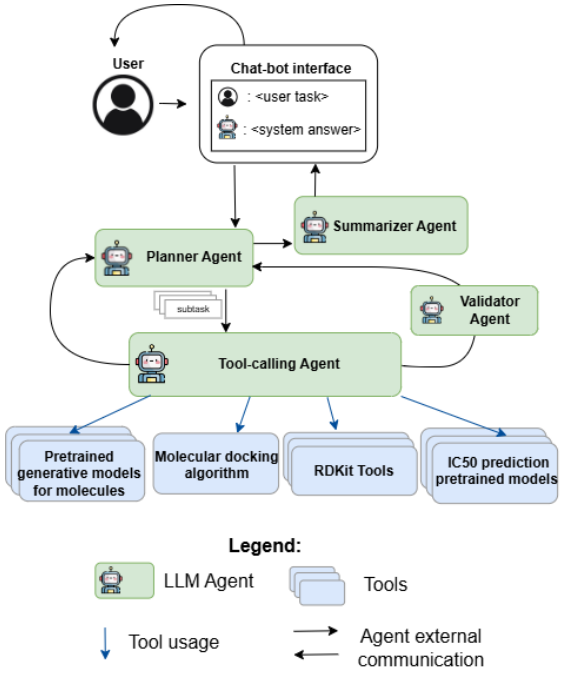

Figure 1: Architecture of the proposed multi-agent approach allowing to decompose a complex drug discovery process into several feasible tasks that can be addressed in an automated way by R&D workers using natural language.

Integrated tools

Our research addresses the complex challenge of real-world drug discovery, which demands a sophisticated integration of multiple specialized tools beyond basic chemical analysis. To meet these requirements, we developed a toolset combining deep generative models for molecules creation, ML models for properties prediction, models for evaluation of synthetic accessibility, drug similarity and other structural properties of molecules.

Generative models. As follows from the related works, the variety of generative models used in drug design is vast. In this work, we refrain from advocating in favor of any particular approach and employ multiple alternatives to molecule generation instead. More specifically, we consider generative adversarial networks (GANs), autoencoders, and reinforcement learning. For those, we adapt existing implementations of LSTM, transformer and graph convolution networks, respectively.

Our GAN implementation consists of 2 LSTM blocks with one bidirectional layer, as well as input layer and hidden layer of size 128. Inspired by the transformer-based conditional VAE (Kim, Na, and Lee 2021), we implemented our own transformer for targeted generation with property control. We trained this model with 8 properties in the conditional block and a vocabulary size of 126 to encode SMILES molecules. The number of transformer layers and heads in the encoder and decoder was also increased to 12. Finally, we extended the recently proposed FREED++ reinforcement learning (RL) framework (Telepov et al. 2024) to also take

into account the 8 target properties while generating molecular candidates.

The model based on FREED++ did not require any additional training as the molecular properties could be estimated and therefore optimized during the generation cycle. In cases of GAN and transformer, training steps were necessary. We pretrained both generative models on the reduced ChEMBL dataset containing 1.2M molecules with molecular weight up to 400 g/mol.

Discriminative models for predicting the activity of generated molecules. Data from ChEMBL and BindingDB were used to create machine learning models for predicting the efficacy of inhibitors of GSK-3, BTK and ABL2. In the case of BTK inhibitors, the data were supplemented from a recent paper (Li et al. 2024a) that also utilises ML for this task. The original data set was presented as molecules in SMILES format and IC50 values (nmol/L). In each case, the necessary data pre-processing was performed in the form of data normalization and duplicate removal. The IC50 prediction task was formulated as a binary classification. The molecules in the data set were divided into two classes by the median of the lgIC50 distribution. Thus, molecules with lgIC50 less than the median were defined as "active" and all others as "inactive". The structures of the molecules were represented in various ways, in particular Morgan fingerprints, Avalon and RDKit descriptors.

Docking score estimation. One of the target properties we used as training data was the binding energy of the target protein to a ligand. This energy can be estimated through molecular docking, typically called the docking score. We calculated docking scores for the disease-specific target proteins using AutoDock Vina (Eberhardt et al. 2021) and QuickVina GPU 2.1 (Tang et al. 2024) frameforks. The latter allowed us to significantly reduce the time required for docking score calculations, averaging just 0.14 seconds compared to 5 seconds with AutoDock Vina. As a result, the total time needed to calculate molecular docking scores for our dataset dropped from 1667 hours to 19 hours.

RDkit-based tools. The Tool-calling Agent may use two RDKit functions: for synthetic accessibility (SA) and drug similarity (QED) estimation. Except this several RDKit-based functions were implemented: such structural filters as Brenk, SurehEMBL, Glaxo, and PAINS.

Experimental setup

Selection criteria for generated molecules

Based on the generation results, the filter was carried out by five stringency groups after calculating all key properties of the molecules. This was done to compare different requirements for compliance with the target properties. Thus, the filtering groups have the following structure:

- **Group 1 (GR1):** *Docking score* $\leq$ *-7 and IC50 = 1*

  This is the main group of filters that considers the biological activity of the generated molecules, the properties of which are proposed to be used as a primary focus.

- **Group 2 (GR2):** *SA score* $\leq$ *3*

  Here, filtering by the possibility of synthesizing substances (SA) to the filters in the first group. This level of filtering additionally shows how many of the generated molecules can potentially be synthesized.

- **Group 3 (GR3):** *Brenk = 0*

  The Brenk filter removes molecules that contain substructures with undesirable pharmacokinetics or toxicity.

- **Group 4 (GR4):** *SureChEMBL = 0*, *Glaxo = 0*, *and PAINS = 0*

  SureChEMBL is a publicly available resource containing compounds extracted from the patent documents. Glaxo filters are designed to exclude unstable and other problematic compound classes. Pan-assay interference compounds (PAINS) are chemical compounds that often give false positive results in high-throughput screens. PAINS tend to react non-specifically with numerous biological targets, which often leads to side effects.

- **Group 5 (GR5):** *QED* $>$ *0.6*

  The most stringent group in terms of filtering includes restrictions on the QED property. Thus, when requiring the inclusion of an assessment of molecules by drug similarity, it is necessary to focus on the fifth group.

Validation dataset preparation

The initial validation dataset, which subsequently served as the basis for generating modified versions for experimental purposes, consists of 245 potential user queries involving mentions of target proteins, properties, and disease symptoms. Examples of these user queries are given in Appendix 6. The dataset was constructed through the following steps:

1. **Initial query design**. 30 queries were manually composed representing theoretical examples that could be posed by the users of different levels of expertise in chemistry. Each query was labeled with a corresponding disease/property name. Most of these queries did not explicitly specify the type of task (e.g., generation/properties calculation) or request invocation of a specific function.
2. **Dataset expansion via few-shot learning**. The dataset was expanded using few-shot learning techniques applied to several LLMs, including *GPT-4o*, *o1-mini*, *Claude Sonnet 3.5*, and *Gemini 1.5 Pro*. The LLMs were provided with a few examples and instructions to generate similar but non-redundant examples. Instructions included explicit requests to generate some examples from a perspective of an experienced professional and a beginner. Upon completion, the dataset was expanded to the total of 400 queries.
3. **Validation by chemistry experts**. Chemistry experts reviewed the synthetically generated queries and selected the most plausible ones. This step yielded the final dataset of 245 queries.
4. **Technical annotation**. The validated dataset underwent annotation by a technical specialist to facilitate downstream processing in validation modules.

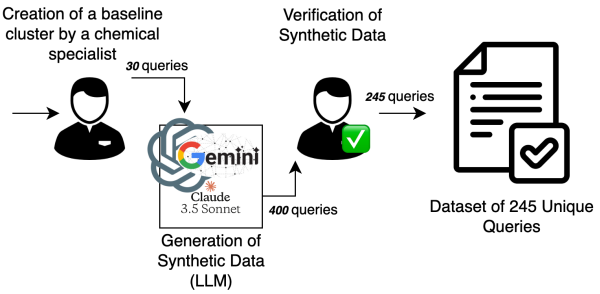

Figure 2: The process of obtaining a validation dataset for further experiments on agents

Examples of user queries are provided in the Appendix A.1.

Experimental conditions

Experiment #1. Comparison of LLMs in the multi-agent system. Experiment 1 focused on analyzing the routing quality for complex queries consisting of up to 5 tasks based on the type of system prompt and the choice of the underlying LLM. The objective was to assess the performance of the pipeline as a whole, utilizing different LLMs and tracking quality metrics for the *Planner Agent* and the *Tool-calling Agent* across queries with varying levels of complexity.

Two prompting strategies were evaluated:

1. Unified system prompt. A universal prompt is applied uniformly across all models.
2. Optimized individual prompts. Tailored prompts were designed to maximize each model's performance. This strategy included two types of prompts: a baseline optimized prompt, enhanced prompt with additional tokens (by including keywords in function descriptions and more precise instructions).

As a metric was used Accuracy (%) of choosing the correct route (was defined in a benchmark) by agents.

Experiment #2. Comparison against a single-agent system. Experiment #2 focused on comparison of routing quality between proposed multi-agent implementation with single-agent implementation. For a clear experiment we use unified core (Langchain based agent) for both implementations. It is worth noting that both implementations share the same set of tools. A individual system prompt was used throughout the experiments, except for the agent-specific instructions. In the single-agent implementation, instructions were updated to align with the single-agent logic. *Llama-3.1-70b* was employed for both implementations as the best performing model.

Two levels of task complexity were evaluated:

1. Requests with 1–3 subtasks. Requests with small amount of details that have to be taken into account by an agentic system.
2. Requests with 4–5 subtasks. More complex and detailed requests.

As a metric was used Accuracy (%) of choosing the correct route (was defined in a benchmark) by agents with differentiation between correct routing on a step of tool calling and routing through the whole task.

Experiment #3. Comparative analysis of generative models efficiency. Experiment #3 is devoted to the demonstration of efficiency of generative models for drug discovery for considered diseases. A key metric for evaluating drug candidate molecules is their ability to meet target properties. We assessed this by calculating the percentage of generated molecules that satisfied our filtering criteria while remaining both novel and chemically valid. This metric was calculated across multiple experimental runs for each model to ensure reliable comparison.

As a metric was used percentage of remained target molecules after filtering with criteria groups.

Experiment #4. Property prediction models validation. Experiment #4 is devoted to evaluation of ML models for property prediction. In order to select the best models for lgIC50 prediction, cross-validation was performed for CatBoost (Prokhorenkova et al. 2018), XGBoost (Chen and Guestrin 2016), Random Forest (Breiman 2001), Extra Trees (Geurts, Ernst, and Wehenkel 2006), and LightGBM (Ke et al. 2017) models. Selection of the best models for each task was made from 5 candidates. Table 5 in Appendix A.4 shows comparative candidate's results.

Experimental results

Demonstration of agentic pipeline efficiency

Experiment #1 results. Comparison of LLMs in the multi-agent system We found that *Llama-3.1-70b* with an optimized system prompt was the best-performing model for routing agents (Figure 3). Notably, this model outperformed the strong baseline of *o1-mini* model and the newer generation *Llama-3.2-90b*, both of which also incurred higher costs. Not unexpectedly, *Llama-3.1-8b* being a smaller and less capable model showed rather poor performance. Pursuing performance and cost efficiency we also included a quantized version of *Llama-3.1-70b* for comparison, which delivered the worst accuracy.

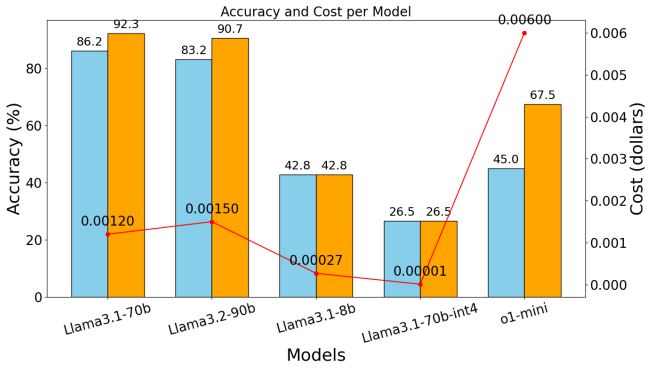

Figure 3: Comparison of the quality and cost of agent pipelines for different models and system prompts.

Experiment #2 results. Comparison against a single-agent system The single-agent system failed to process any requests with 4–5 tasks. While functions were correctly identified and invoked, the final responses were not valid (e.g. include no generated molecules, contained fewer than half of the expected outputs, etc.).

For smaller requests (1-3 subtasks), the single-agent system produced better results but still fell short compared to the multi-agent implementation. The multi-agent system provides a clear, structured response by separating molecules by task type and ensuring completeness. In contrast, a single-agent system often misses important properties or confounds unrelated results.

In particular, on queries consisting of 1-3 tasks, the multi-agent system achieved 9% higher accuracy in function calls and 20% higher precision in the whole pipeline, as shown in the Table 1. For queries of 4-5 tasks, based on the inability of the single-agent system to produce a final complete answer, its accuracy was 0%. At the same time, the single-agent system did not do so poorly in function detection, yielding to the multi-agent system by 18%.

Table 1: The accuracy of the entire pipeline and function calls in both multi-agent and single-agent implementations for different number of subtasks in user query.

Number of tasks	Pipeline	Accuracy	
		Function call	Overall pipeline
1-3	Multi-Agent	**0.92**	**0.92**
	Single-Agent	0.82	0.71
4-5	Multi-Agent	**0.92**	**0.92**
	Single-Agent	0.74	0.0

Demonstration of designed case study practical valuability

Beyond validating the agentic pipeline efficiency, we provide additional experimental evidence demonstrating our approach practical value for state-of-the-art drug discovery research. Since our goal is to advance automated pharmaceutical design rather than solve textbook chemistry problems, we focused on complex therapeutic targets in neurodegenerative diseases.

Experiment #3. Comparative analysis of generative models efficiency. The key finding from Experiment #3 is a demonstration of ability to generate valid novel molecules that pass multiple pharmaceutical filters (with non-zero success rate). Both GAN and Transformer based models successfully generated viable drug candidates across all three disease targets. The FREED++ model proved effective for two targets (Alzheimer's and Multiple Sclerosis) but failed to generate valid molecules for Parkinson's disease. Comparison results are presented in Table 2.

Another important property for generative models inside drug design benchmark is an ability to create diverse molecules that is estimated using Tanimoto similarity ((Bajusz, Rácz, and Héberger 2015)). As shown in Table 2 (and Figure 5 in Appendix A.4) all models give a reasonable level of diversity. Although, GAN-based models produced more diverse molecules compared to Transformer and RL approaches. While the latter models prioritized target properties over diversity, leading to higher success rates in generating viable drug candidates.

Also following (Telepov et al. 2024), we analyzed the top 25 molecules from each model (Table 4 in Appendix A.3). All molecules achieved zero scores on key drug-likeness filters (Brenk, SureChEMBL, Glaxo, PAINS).

Experiment #4. Property prediction models validation. Selected results of the best chosen models are shown in Table 3. Such results suggest that molecules potentially applicable for the treatment of a particular disease and have specific combinations of properties. However, the results for multiple sclerosis show a slight decrease in accuracy compared to a recent study (Li et al. 2024a) for predicting the activity of BTK inhibitors (5.32%), which is probably due to an extended training dataset in our work and the resulting improved generalization ability.

Detailed evaluation: Alzheimer's disease In addition to evaluating individual molecular properties, we conducted comprehensive analysis of the drug candidates generated through our benchmark. Due to limited volume, we present here detailed validation results only for the Alzheimer's disease case.

16,082 novel GSK-3β inhibitors were generated using the transformer model. To validate generated molecules with already known compounds, we compared novel with active inhibitors from the ChEMBL dataset, which was used to create the IC50 prediction model (see Figure 4). The average SA Score of the generated molecules is lower than experimentally validated compounds, suggesting easier ways of laboratory synthesis. Moreover, the average QED score of generated molecules increased 11.8%, which indicated enhanced pharmacological properties. Lower toxicity can also be reported since all the generated molecules have passed the Brenk filter. At the same time, the Tanimoto similarity of 0.43 between novel and ChEMBL molecules leads to a conclusion that along with improved properties, the obtained compounds make up a different chemical space, which can potentially result in unconventional and effective solutions for this case (Ganeeva et al. 2024).

Conclusion and Discussion

In the paper, we aimed to demonstrate that transitioning from traditional human-in-ML-loop drug discovery to fully automated drug discovery is more challenging than commonly assumed, particularly when moving beyond textbook chemistry problems to state-of-the-art pharmaceutical research tasks. This conclusion is supported by *two key experimental findings*.

First, we demonstrate that automating the transition from unstructured text queries to valid molecular structures requires multi-agent approach. While single-agent LLM implementations perform adequately on simple queries (71% accuracy), they fail on complex drug discovery tasks that require coordinating multiple specialized tools and extended

Table 2: Percentage of target molecules across filter groups obtained during the generation series by each model.

Case	Model	GR1, %	GR2, %	GR3, %	GR4, %	GR5, %	Diversity
Alzheimer	GAN	19.03	14.75	11.70	11.32	11.32	**0.37**
	Transformer	**26.06**	**23.58**	**18.47**	**18.15**	**18.15**	0.24
	RL	15.8	14.34	10.99	10.74	10.74	0.21
Multiple sclerosis	GAN	5.90	4.35	3.49	3.36	3.36	**0.39**
	Transformer	15.43	13.75	13.32	13.29	13.29	0.25
	RL	**22.81**	**20.34**	**18.39**	**18.22**	**18.22**	0.11
Parkinson	GAN	**14.45**	**11.48**	**8.92**	**8.57**	**8.57**	**0.36**
	Transformer	3.32	3.06	2.69	2.65	2.65	0.24
	RL	0.03	0.03	0	0	0	0.17

Table 3: Performance of machine learning models predicting activity of the generated molecules (binary classification based on the case-specific lgIC50 threshold).

Case	The best model	Accuracy	F1-score
Alzheimer	Extra Trees	0.82	0.83
Multiple sclerosis	Random Forest	0.89	0.92
Parkinson	Catboost	0.91	0.92

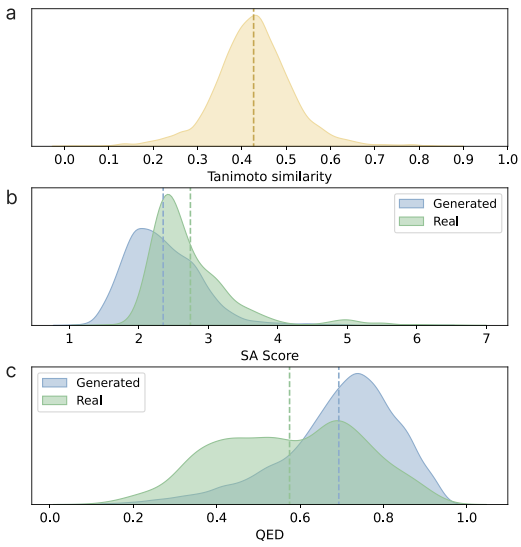

Figure 4: Complex validation of generative pipeline for Alzheimer's disease case: a) Distribution of maximum Tanimoto Similarity between generated molecules and experimentally validated GSK-3 inhibitors; b) SA Score distributions, c) QED distributions. "Generated" are molecules generated using generative models, "Real" are experimentally validated GSK-3 inhibitors

planning. Proposed multi-agent system achieves 92% accuracy by effectively decomposing and planning routes for complex tasks solution.

Despite the function calls achieving an accuracy of 74%, gathering information on all tasks and formulating a final response for the user were not carried out correctly, resulting in zero accuracy, in contrast to the multi-agent pipeline.

Second, we demonstrate that our pipeline can address real pharmaceutical challenges rather than just academic examples through experiments on brain degenerative disease cases. While implemented queries, generative models, property prediction models altogether may form a new benchmark with comprehensive pharmaceutical research pipeline for solution of current scientific tasks.

Our results suggest that while LLM-based automation of drug discovery shows promise, the gap between automated and traditional approaches remains significant. The multi-agent approach provides a viable path forward, but requires careful consideration of both computational and domain-specific challenges. This highlights the importance of rigorous validation against real pharmaceutical research problems rather than relying solely on simplified academic examples.

It appears that single-LLM approaches currently significantly lag behind in capabilities compared to results achievable through agent-based approaches that leverage a wide range of tools and ML models. Our experiments showed that general models like *Llama 3.1*, *GPT-4o*, *o1-mini* fail to generate valid molecules in the vast majority of cases. However, specialized LLM models face similar limitations: for instance, *ChemLLM* cannot solve the molecule generation task at all, while *DrugLLM* is designed only for optimizing existing formulas. While single-LLM approach capabilities undoubtedly require more detailed quantitative analysis and comprehensive comparison with agent-based approaches, that remains as future work.

Besides all, objective comparison between multi-agent and single-agent implementations remains an open discussion point, involving numerous architectural and implementation nuances (prompt selection, multi-agent network structure, choice of base frameworks, etc.). While additional research is required for definitive conclusions, we believe our experimental results support the validity of our findings at the current level of tested hypotheses.

Another important future work is extending the existing test cases beyond brain diseases to a new comprehensive benchmark that includes possible queries, candidate molecules, filtered through criteria set molecules and final answers.

References

Allenspach, S.; Hiss, J. A.; and Schneider, G. 2024. Neural multi-task learning in drug design. *Nature Machine Intelligence*, 6(2): 124–137.

Ang, D.; Rakovski, C.; and Atamian, H. S. 2024. De Novo Drug Design Using Transformer-Based Machine Translation and Reinforcement Learning of an Adaptive Monte Carlo Tree Search. *Pharmaceuticals*, 17(2): 161.

Bajusz, D.; Rácz, A.; and Héberger, K. 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7: 1–13.

Bhadwal, A. S.; Kumar, K.; and Kumar, N. 2023. GMG-NCDVAE: guided de novo molecule generation using NLP techniques and constrained diverse variational autoencoder. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.

Buerger, K.; Ewers, M.; Pirttilä, T.; Zinkowski, R.; Alafuzoff, I.; Teipel, S. J.; DeBernardis, J.; Kerkman, D.; McCulloch, C.; Soininen, H.; et al. 2006. CSF phosphorylated tau protein correlates with neocortical neurofibrillary pathology in Alzheimer's disease. *Brain*, 129(11): 3035–3041.

Cencioni, M. T.; Mattoscio, M.; Magliozzi, R.; Bar-Or, A.; and Muraro, P. A. 2021. B cells in multiple sclerosis—from targeted depletion to immune reconstitution therapies. *Nature Reviews Neurology*, 17(7): 399–414.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Doherty, T.; Yao, Z.; Khleifat, A. A.; Tantiangco, H.; Tamburin, S.; Albertyn, C.; Thakur, L.; Llewellyn, D. J.; Oxtoby, N. P.; Lourida, I.; et al. 2023. Artificial intelligence for dementia drug discovery and trials optimization. *Alzheimer's & Dementia*, 19(12): 5922–5933.

Dollar, O.; Joshi, N.; Beck, D. A.; and Pfaendtner, J. 2021. Attention-based generative models for de novo molecular design. *Chemical Science*, 12(24): 8362–8372.

Domínguez, J. M.; Fuertes, A.; Orozco, L.; del Monte-Millán, M.; Delgado, E.; and Medina, M. 2012. Evidence for irreversible inhibition of glycogen synthase kinase-3β by tideglusib. *Journal of Biological Chemistry*, 287(2): 893–904.

Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; and Forli, S. 2021. AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8): 3891–3898.

Ganeeva, V.; Sakhovskiy, A.; Khrabrov, K.; Savchenko, A.; Kadurin, A.; and Tutubalina, E. 2024. Lost in Translation: Chemical Language Models and the Misunderstanding of Molecule Structures. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12994–13013.

Garg, V. 2024. Generative AI for graph-based drug design: Recent advances and the way forward. *Current Opinion in Structural Biology*, 84: 102769.

Geurts, P.; Ernst, D.; and Wehenkel, L. 2006. Extremely randomized trees. *Machine learning*, 63: 3–42.

Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; and Aspuru-Guzik, A. 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2): 268–276.

Grisoni, F. 2023. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology*, 79: 102527.

Grisoni, F.; Moret, M.; Lingwood, R.; and Schneider, G. 2020. Bidirectional molecule generation with recurrent neural networks. *Journal of chemical information and modeling*, 60(3): 1175–1183.

Guan, S.; and Wang, G. 2024. Drug discovery and development in the era of artificial intelligence: From machine learning to large language models. *Artificial Intelligence Chemistry*, 2(1): 100070.

Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; and Aspuru-Guzik, A. 2017. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*.

Haroon, S.; Hafsath, C.; and Jereesh, A. 2023. Generative pre-trained transformer (GPT) based model with relative attention for de novo drug design. *Computational Biology and Chemistry*, 106: 107911.

Huang, D.; Yang, M.; Wen, X.; Xia, S.; and Yuan, B. 2024. AI-driven drug discovery:: accelerating the development of novel therapeutics in biopharmaceuticals. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, 3(3): 206–224.

Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; et al. 2023. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5): 1233–1250.

Jensen, J. H. 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.

Jha, N. K.; Chen, W.-C.; Kumar, S.; Dubey, R.; Tsai, L.-W.; Kar, R.; Jha, S. K.; Gupta, P. K.; Sharma, A.; Gundamaraju, R.; et al. 2022. Molecular mechanisms of developmental pathways in neurological disorders: a pharmacological and therapeutic review. *Open biology*, 12(3): 210289.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Kerstjens, A.; and De Winter, H. 2022. LEADD: Lamarckian evolutionary algorithm for de novo drug design. *Journal of Cheminformatics*, 14(1): 3.

Khachaturian, A.; Dengel, A.; Dočkal, V.; Hroboň, P.; and Tolar, M. 2023. Accelerating Innovations for Enhanced Brain Health. Can Artificial Intelligence Advance New Pathways for Drug Discovery for Alzheimer's and other Neurodegenerative Disorders? *The Journal of Prevention of Alzheimer's Disease*, 10(1): 1–4.

Kim, H.; Na, J.; and Lee, W. B. 2021. Generative chemical transformer: neural machine learning of molecular geometric structures from chemical language via attention. *Journal of chemical information and modeling*, 61(12): 5804–5814.

Krämer, J.; Bar-Or, A.; Turner, T. J.; and Wiendl, H. 2023. Bruton tyrosine kinase inhibitors for multiple sclerosis. *Nature Reviews Neurology*, 19(5): 289–304.

Kwon, S.-H.; Kim, S.; Park, A. Y.; Lee, S.; Gadhe, C. G.; Seo, B. A.; Park, J.-S.; Jo, S.; Oh, Y.; Kweon, S. H.; et al. 2021. A novel, selective c-Abl inhibitor, compound 5, prevents neurodegeneration in Parkinson's disease. *Journal of medicinal chemistry*, 64(20): 15091–15110.

Latif, E.; Parasuraman, R.; and Zhai, X. 2024. PhysicsAssistant: An LLM-Powered Interactive Learning Robot for Physics Lab Investigations. *arXiv preprint arXiv:2403.18721*.

Lee, M.; and Min, K. 2022. MGCVAE: multi-objective inverse design via molecular graph conditional variational autoencoder. *Journal of chemical information and modeling*, 62(12): 2943–2950.

Leguy, J.; Cauchy, T.; Glavatskikh, M.; Duval, B.; and Da Mota, B. 2020. EvoMol: a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation. *Journal of cheminformatics*, 12: 1–19.

Li, G.; Li, J.; Tian, Y.; Zhao, Y.; Pang, X.; and Yan, A. 2024a. Machine learning-based classification models for non-covalent Bruton's tyrosine kinase inhibitors: Predictive ability and interpretability. *Molecular Diversity*, 28(4): 2429–2447.

Li, R.; Tang, H.; Burns, J. C.; Hopkins, B. T.; Le Coz, C.; Zhang, B.; de Barcelos, I. P.; Romberg, N.; Goldstein, A. C.; Banwell, B. L.; et al. 2022. BTK inhibition limits B-cell–T-cell interaction through modulation of B-cell metabolism: implications for multiple sclerosis therapy. *Acta neuropathologica*, 143(4): 505–521.

Li, T.; Shetty, S.; Kamath, A.; Jaiswal, A.; Jiang, X.; Ding, Y.; and Kim, Y. 2024b. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *NPJ Digital Medicine*, 7(1): 40.

Li, X.; Xu, Y.; Yao, H.; and Lin, K. 2020. Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. *Journal of cheminformatics*, 12: 1–13.

Liu, X.; Guo, Y.; Li, H.; Liu, J.; Huang, S.; Ke, B.; and Lv, J. 2024. DrugLLM: Open Large Language Model for Few-shot Molecule Generation. *arXiv preprint arXiv:2405.06690*.

M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; and Schwaller, P. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 1–11.

Macedo, B.; Ribeiro Vaz, I.; and Taveira Gomes, T. 2024. MedGAN: optimized generative adversarial network with graph convolutional networks for novel molecule design. *Scientific Reports*, 14(1): 1212.

Mao, J.; Wang, J.; Zeb, A.; Cho, K.-H.; Jin, H.; Kim, J.; Lee, O.; Wang, Y.; and No, K. T. 2023. Transformer-based molecular generative model for antiviral drug design. *Journal of chemical information and modeling*, 64(7): 2733–2745.

McGinley, M. P.; Goldschmidt, C. H.; and Rae-Grant, A. D. 2021. Diagnosis and treatment of multiple sclerosis: a review. *Jama*, 325(8): 765–779.

McNaughton, A. D.; Sankar Ramalaxmi, G. K.; Kruel, A.; Knutson, C. R.; Varikoti, R. A.; and Kumar, N. 2024. CACTUS: Chemistry Agent Connecting Tool Usage to Science. *ACS Omega*.

Ochiai, T.; Inukai, T.; Akiyama, M.; Furui, K.; Ohue, M.; Matsumori, N.; Inuki, S.; Uesugi, M.; Sunazuka, T.; Kikuchi, K.; et al. 2023. Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity. *Communications Chemistry*, 6(1): 249.

Pang, C.; Qiao, J.; Zeng, X.; Zou, Q.; and Wei, L. 2023. Deep generative models in de novo drug molecule generation. *Journal of Chemical Information and Modeling*, 64(7): 2174–2194.

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; and Chen, H. 2019. A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics*, 11: 1–13.

Pushkaran, A. C.; and Arabi, A. A. 2024. From understanding diseases to drug design: can artificial intelligence bridge the gap? *Artificial Intelligence Review*, 57(4): 86.

Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; and Zhavoronkov, A. 2018. Reinforced adversarial neural computer for de novo molecular design. *Journal of chemical information and modeling*, 58(6): 1194–1204.

Saberi, D.; Geladaris, A.; Dybowski, S.; and Weber, M. S. 2023. Bruton's tyrosine kinase as a promising therapeutic target for multiple sclerosis. *Expert Opinion on Therapeutic Targets*, 27(4-5): 347–359.

Suresh, N.; Chinnakonda Ashok Kumar, N.; Subramanian, S.; and Srinivasa, G. 2022. Memory augmented recurrent neural networks for de-novo drug design. *Plos one*, 17(6): e0269461.

Tang, S.; Ding, J.; Zhu, X.; Wang, Z.; Zhao, H.; and Wu, J. 2024. Vina-GPU 2.1: towards further optimizing docking speed and precision of AutoDock Vina and its derivatives. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Telepov, A.; Tsypin, A.; Khrabrov, K.; Yakukhnov, S.; Strashnov, P.; Zhilyaev, P.; Rumiantsev, E.; Ezhov, D.;

Avetisian, M.; Popova, O.; and Kadurin, A. 2024. FREED++: Improving RL Agents for Fragment-Based Molecule Generation by Thorough Reproduction. Version Number: 1.

Thomas, M.; O'Boyle, N. M.; Bender, A.; and De Graaf, C. 2022. Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. *Journal of cheminformatics*, 14(1): 68.

Tolosa, E.; Garrido, A.; Scholz, S. W.; and Poewe, W. 2021. Challenges in the diagnosis of Parkinson's disease. *The Lancet Neurology*, 20(5): 385–397.

Tripp, A.; and Hernández-Lobato, J. M. 2023. Genetic algorithms are strong baselines for molecule generation. *arXiv preprint arXiv:2310.09267*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need.(Nips), 2017. *arXiv preprint arXiv:1706.03762*, 10: S0140525X16001837.

Vicidomini, C.; Fontanella, F.; D'Alessandro, T.; and Roviello, G. N. 2024. A Survey on Computational Methods in Drug Discovery for Neurodegenerative Diseases. *Biomolecules*, 14(10): 1330.

Werner, M. H.; and Olanow, C. W. 2022. Parkinson's disease modification through Abl kinase inhibition: an opportunity. *Movement Disorders*, 37(1): 6–15.

Wu, D.; Chen, Q.; Chen, X.; Han, F.; Chen, Z.; and Wang, Y. 2023. The blood–brain barrier: structure, regulation, and drug delivery. *Signal Transduction and Targeted Therapy*, 8(1): 217.

Ye, G. 2024. De novo drug design as GPT language modeling: large chemistry models with supervised and reinforcement learning. *Journal of Computer-Aided Molecular Design*, 38(1): 20.

Yoshikawa, N.; Terayama, K.; Sumita, M.; Homma, T.; Oono, K.; and Tsuda, K. 2018. Population-based de novo molecule generation, using grammatical evolution. *Chemistry Letters*, 47(11): 1431–1434.

Zeng, X.; Wang, F.; Luo, Y.; Kang, S.-g.; Tang, J.; Lightstone, F. C.; Fang, E. F.; Cornell, W.; Nussinov, R.; and Cheng, F. 2022. Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine*, 3(12).

Zhang, D.; Liu, W.; Tan, Q.; Chen, J.; Yan, H.; Yan, Y.; Li, J.; Huang, W.; Yue, X.; Zhou, D.; et al. 2024a. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Zhang, Z.; Bo, X.; Ma, C.; Li, R.; Chen, X.; Dai, Q.; Zhu, J.; Dong, Z.; and Wen, J.-R. 2024b. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

Zhao, Z.; Ma, D.; Chen, L.; Sun, L.; Li, Z.; Xu, H.; Zhu, Z.; Zhu, S.; Fan, S.; Shen, G.; et al. 2024. Chemdfm: Dialogue foundation model for chemistry. *arXiv preprint arXiv:2401.14818*.

Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; et al. 2019. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology*, 37(9): 1038–1040.

Appendix

A.1 - Examples of Queries:

Manually Composed Queries by Experts

- **Alzheimer's Disease:**
  - *Generate GSK-3beta inhibitors with high docking score and low brain-blood barrier permeability.*
  - *Generate GSK-3beta inhibitors with high activity.*

Synthetic Queries Generated by LLMs

- **Alzheimer's Disease:**
  - *Generate structures with many aromatic rings to facilitate π-π stacking interactions with beta-amyloid aggregates.*
  - *Generate compounds with flexible linkers to allow for conformational adaptability in binding to amyloid-beta.*

A.2 - Examples of single-agent responses:

- **Response 1:** *"Here are the generated GSK-3beta inhibitors with high activity for the treatment of Alzheimer"* (no molecules provided).
- **Response 2:** *"Here are the generated GSK-3beta inhibitors..."* (identical molecules were duplicated for each task, ignoring valid outputs and missing properties).

A.3 - Properties of generated molecules:

Table 4: Average values of key properties for top-25 generated molecules

Case	Models	Avg Docking score	Avg SA	Avg QED
Alzheimer	GAN	-9.5	2.4	0.49
	RL (FREED++)	-9.1	**2.3**	**0.54**
	Transformer	**-10.7**	2.5	0.51
Mupltiple sklerosis	GAN	-11.4	2.3	0.73
	RL (FREED++)	-11.4	**2.3**	**0.74**
	Transformer	**-11.5**	2.5	0.65
Parkinson	GAN	**-9.3**	2.3	0.40
	RL (FREED++)	-6.1	2.25	0.51
	Transformer	-8.2	**2.1**	**0.60**

A.4 - Additional ML results:

The more detailed results for ML experiments are provided in Table 5 and Figure 5.

A.5 - Analysed brain disease cases:

Alzheimer's disease Currently, there are no medications that fully prevent or halt Alzheimer's disease (AD). Existing drugs only reduce symptoms. Tau proteins play a role in stabilizing microtubules, which maintain the healthy state of neurons (Buerger et al. 2006). In a healthy brain, tau proteins undergo phosphorylation and dephosphorylation, processes regulated by various kinases. Glycogen synthase kinase-3 (GSK-3) is a serine/threonine kinase that plays a key role in cellular metabolism and signal transduction. It is associated with various diseases, including AD, by promoting tau protein hyperphosphorylation, which is a major component of neurofibrillary tangles, one of the hallmarks of AD. One of the inhibitors of this kinase, tideglusib, has completed phase I and II clinical trials, during which it was found that cognitive function in patients improved slightly compared to placebo (insufficient efficacy), and gastrointestinal side effects (toxicity) were observed (Domínguez et al. 2012). Thus, development of novel GSK-3 inhibitors with enhanced properties is of great importance.

Multiple sclerosis Multiple sclerosis (MS) is a chronic autoimmune disorder affecting the central nervous system, characterized by inflammation, demyelination, gliosis, and neuroaxonal degeneration (McGinley, Goldschmidt, and Rae-Grant 2021). While it is traditionally thought that MS is primarily mediated by T-cells, B-cells and almost all types of innate immune cells appear to play a significant role in both the initiation and propagation of the disease. Peripheral immune cells that cross the blood-brain barrier (BBB) induce relapses and the formation of focal demyelinating plaques (Cencioni et al. 2021). Bruton's tyrosine kinase (BTK) is a protein that plays a critical role in the development and function of immune cells. The use of BTK

Table 5: Comparison of Accuracy and F1 score for the considered machine learning models. For Alzheimer disease case MACCS fingerprints were used, for mupltiple sclerosis - Morgan fingerprints (1024, radius=2), for Parkinson disease - RDKit descriptors and Avalon fingerprints.

Case	Model	Accuracy	F1 Score
Alzheimer	CatBoost	0.810	0.810
	Random Forest	0.822	0.829
	XGBoost	0.803	0.803
	Extra Trees	**0.823**	**0.829**
	LightGBM	0.810	0.820
Mupltiple sclerosis	CatBoost	0.865	0.905
	Random Forest	**0.887**	**0.920**
	XGBoost	0.876	0.912
	Extra Trees	0.886	0.919
	LightGBM	0.885	0.918
Parkinson	**CatBoost**	**0.910**	**0.920**
	Random Forest	0.890	0.900
	XGBoost	0.910	0.910
	Extra Trees	0.890	0.900
	LightGBM	0.900	0.910

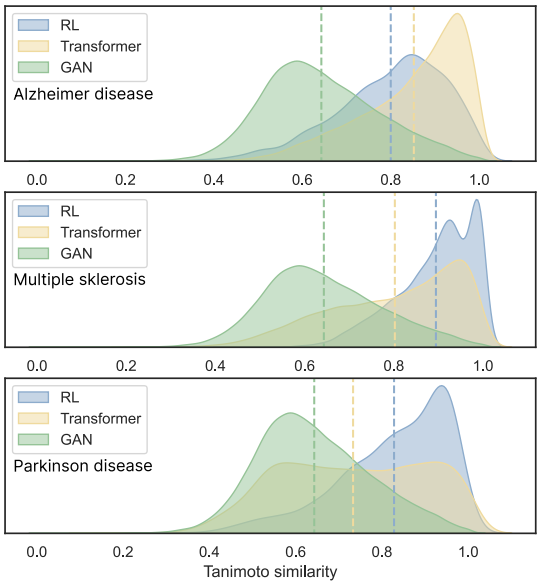

Figure 5: Tanimoto similarity (maximum values) for all generated molecules

inhibitors for the treatment of MS is a promising area of research, as these drugs have been shown to reduce B-cell activity and decrease inflammation in the brain and spinal cord (Krämer et al. 2023). By targeting BTK, these drugs may slow or halt the progression of MS, improve symptoms, and enhance the quality of life of patients (Li et al. 2022). Currently, at least six BTK inhibitors (BIIB091, Evobrutinib, Fenebrutinib, Orelabrutinib, Remibrutinib, Tolebrutinib) are in phase II-III clinical trials. Despite promising results, there are still areas for improvement in BTK inhibitors, such as binding mechanism (non-covalent inhibitors are less potent and require higher doses, but they offer increased selectivity and a lower propensity for resistanc) and blood-brain barrier permeability (Saberi et al. 2023). The objective of this case is to generate noncovalent BTK inhibitors with improved IC50 values and enhanced BBB permeability.

Parkinson's disease Parkinson's disease is a progressive neurodegenerative disorder, which is characterized by the loss of dopaminergic neurons (Tolosa et al. 2021). The primary causes and mechanisms of development include mitochondrial dys-

function, oxidative stress, genetic mutations, protein manifolding and aggregation, as well as disruptions in cellular clearance processes. These factors contribute to neuronal degeneration and make them key targets for therapeutic approaches. Two main targets are being investigated for the treatment of Parkinson's disease: tyrosine protein kinase ABL and catecholamines. Inhibition of ABL is considered a promising approach to slowing neurodegenerative processes (Werner and Olanow 2022). This protein kinase is involved in the regulation of cellular metabolism, and its hyperactivation is associated with increased oxidative stress and the accumulation of damaged proteins, which contribute to neuronal death (Kwon et al. 2021). This case study focuses on the generation of new ABL inhibitors with improved properties.

A.6 - Examples of requests.

Simple requests. Requests with 1–3 subtasks:

- *"Generate GSK-3beta inhibitors with high activity".*
- *"Give me active molecules against GSK-3beta protein. I want them to be not toxic"*
- *"Design one unique molecular entity with a binding affinity greater than 100 nM that specifically targets amyloid-beta plaques associated with Alzheimer's disease. This compound should demonstrate a permeability coefficient of at least 1.5 to ensure effective crossing of the blood-brain barrier."*
- *"Generate inhibitors of KRAS protein with G12C mutation. The inhibitors should be selective, meaning they should not bind with HRAS and NRAS proteins."*
- *"Generate highly potent non-covalent BTK tyrosine kinase inhibitors from the TEC family of tyrosine kinases that have the potential to affect B cells as a therapeutic target for the treatment of multiple sclerosis"*
- *"Generate molecules with activity against Parkinson's disease based on phenethylamine scaffolds. "*
- *"Synthesize dual-action agents that modulate both HDL and LDL particle size for improved cardiovascular outcomes."*
- *"Design irreversible inhibitors of cholinesterases with improved selectivity to enhance cholinergic transmission in the Parkinson's patient population."*
- *"Create one detailed SMILES representation for a potent inhibitor that effectively modulates the activity of ABC transporters implicated in drug resistance. "*
- *"Develop one novel therapeutic candidate that functions as a selective antagonist of the N-methyl-D-aspartate (NMDA) receptor with a Ki value lower than 30 nM. This candidate should prevent excitotoxicity while maintaining synaptic function, showing a balance in action with a minimal effect on synaptic transmission at therapeutic doses. Design novel small molecules targeting the efflux pumps responsible for drug resistance."*
- *"Generate derivatives that incorporate multi-targeted inhibition to address the complex mechanisms underlying Alzheimer's disease.Create 1 selective molecule that targets KRAS G12C and doesn't affect HRAS or NRAS. Generate molecular structures targeting drug resistance mechanisms in cancer cells."*
- *"Synthesize a potent and selective ANGPTL3 inhibitor to reduce plasma triglycerides."*
- *"Generate molecules with inhibitory activity against glycogen synthase kinase 3 beta (GSK-3) to reduce tau phosphorylation in Alzheimer's pathology. Create 1 selective molecule that targets KRAS G12C and doesn't affect HRAS or NRAS. Generate molecular structures targeting drug resistance mechanisms in cancer cells."*

Complex requests. Requests with 4–5 subtasks:

- *"Generate GSK-3beta inhibitors with high activit. Suggest some small molecules that inhibit KRAS G12C - a target responsible for non-small cell lung cancer. Generate high activity tyrosine-protein kinase BTK inhibitors. Generate me 2 molecules that would help me with my blood lipid spectrum disorder, which is manifested by an increase in cholesterol, triglycerides, low and very low density lipoproteins and a decrease in high density lipoproteins, or alpha lipoproteins. It is important that medications do not produce side effects such as muscle pain and liver problems."*
- *"Suggest several molecules that have high docking affinity with KRAS G12C protein. Molecules should possess common drug-like properties, including low toxicity, high QED score, and high level of synthesizability. Generate high activity tyrosine-protein kinase BTK inhibitors. Can you suggest molecules that inhibit Proprotein Convertase Subtilisin/Kexin Type 9 with enhanced bioavailability and the ability to cross the BBB? Generate me new drug that enhance neurotransmitter balance, promote neuroprotection, and reduce oxidative stress. These compounds should possess high bioavailability, cross the blood-brain barrier efficiently, and show minimal metabolic degradation."*
- *"Synthesize compounds that inhibit the phosphoinositide 3-kinase pathway in resistant cancers. Generate molecules with inhibitory activity against glycogen synthase kinase 3 beta (GSK-3) to reduce tau phosphorylation in Alzheimer's pathology. Suggest some small molecules that inhibit KRAS G12C - a target responsible for non-small cell lung cancer. Generate highly potent non-covalent BTK inhibitors that will have increased permeability through the blood-brain barrier. Generate me 2 molecules that would help me with my blood lipid spectrum disorder, which is manifested by an increase in cholesterol, triglycerides, low and very low density lipoproteins and a decrease in high density lipoproteins, or alpha lipoproteins. It is important that medications do not produce side effects such as muscle pain and liver problems."*

- *"Discover therapeutic agents targeting non-covalent BTK modulation to prevent multiple sclerosis progression. Create novel small molecules to specifically bind and inhibit KRAS G12C, ensuring no activity against HRAS and NRAS. Generate me 2 molecules that could help in the treatment of Parkinson's disease, focusing on compounds that support the regulation of dopamine levels and protect neurons from oxidative stress and mitochondrial dysfunction. It is important that these molecules do not cause severe side effects such as hallucinations, dyskinesia, or cardiovascular issues. Generate me 2 molecules that could overcome chemotherapeutic resistance in cancer treatment, specifically targeting mechanisms such as increased drug efflux, enhanced DNA repair, or apoptosis evasion. It is important that these compounds avoid toxicity to healthy cells and minimize side effects like immunosuppression or gastrointestinal distress."*
- *"Design 3 novel inhibitors targeting KRAS G12C for lung cancer treatment, ensuring high selectivity and no binding to HRAS or NRAS. Generate highly potent non-covalent BTK inhibitors that will have increased permeability through the blood-brain barrier. Generate inhibitors of SIRT1 to modulate lipid metabolism and improve insulin sensitivity. Generate molecules with properties of glutamate receptor antagonists for neuroprotection."*
- *"Identify novel small molecules that suppress BTK-mediated pathways, reducing inflammation in multiple sclerosis. Suggest some small molecules that inhibit KRAS G12C - a target responsible for non-small cell lung cancer. Generate me 2 molecules that could help in the treatment of Parkinson's disease, focusing on compounds that support the regulation of dopamine levels and protect neurons from oxidative stress and mitochondrial dysfunction. It is important that these molecules do not cause severe side effects such as hallucinations, dyskinesia, or cardiovascular issues. Generate one synergistic compound that significantly enhances the activity of existing therapeutic agents against drug-resistant pathogens."*
- *"Create selective drug-like inhibitors that target the KRAS G12C mutation in lung cancer, while avoiding off-target activity with HRAS and NRAS proteins. Focus on selectivity for KRAS and avoid off-target effects with other RAS family proteins. Devise novel compounds that interfere with immune signaling pathways to treat multiple sclerosis. Can you suggest molecules that inhibit Proprotein Convertase Subtilisin/Kexin Type 9 with enhanced bioavailability and the ability to cross the BBB? Generate me new drug that enhance neurotransmitter balance, promote neuroprotection, and reduce oxidative stress. These compounds should possess high bioavailability, cross the blood-brain barrier efficiently, and show minimal metabolic degradation."*
- *"Develop selective tyrosine kinase inhibitors with strong binding affinity for BTK. Create novel small molecules to specifically bind and inhibit KRAS G12C, ensuring no activity against HRAS and NRAS. Generate me 2 molecules that could help in the treatment of Parkinson's disease, focusing on compounds that support the regulation of dopamine levels and protect neurons from oxidative stress and mitochondrial dysfunction. It is important that these molecules do not cause severe side effects such as hallucinations, dyskinesia, or cardiovascular issues. Can you suggest molecules that inhibit signal transducer and activator of transcription 3 (STAT3) with water solubility greater than 60 g/mL and inhibitory ability to P450 CYP1A2?"*