LLM Drug Discovery Challenge: A Contest as a Feasibility Study on the Utilization of Large Language Models in Medicinal Chemistry

Kusuri Murakumo
The University of Tokyo
SHaLX Inc.
souyakuchan@shalx.co.jp

Naruki Yoshikawa
University of Toronto

Kentaro Rikimaru
ExaWizards Inc.

Shogo Nakamura, Kairi Furui, Takamasa Suzuki, Masahito Ohue
Tokyo Institute of Technology

Hiroyuki Yamasaki
Cykinso, Inc.

Yuki Nishigaya
AgroDesign Studios Co. Ltd.

Yuzo Takagi
iSiP, Inc.

Abstract

The ultimate ideal in AI-driven drug discovery is the automatic design of specific drugs for individual diseases, yet this goal remains technically distant at present. However, recent advancements in large language models (LLMs) have significantly broadened the scope of applications with various tasks being explored in the chemistry domain. To probe the potential of utilizing LLMs in drug discovery, we organized a contest: the LLM Drug Discovery Challenge. Participants were tasked with proposing molecular structures of active compound candidates for a designated drug target using LLM-based workflows. The proposed chemical structures were evaluated comprehensively through scoring by a panel of five judges with deep expertise in medicinal chemistry, structural biology, and computational chemistry. Nine participants tackled the challenge with their unique methodologies, exploring the possibilities and current limitations of leveraging LLMs in drug discovery. In this rapidly advancing field, we aim to discuss the directions of future developments and what is expected moving forward.

1 Introduction

The quest to automate drug discovery through artificial intelligence (AI) has captivated the scientific community [1, 2], with the ultimate goal of autonomously designing specific drugs for individual diseases. However, despite significant advances in AI research, this goal remains technically distant [3]. In particular, the range of tasks that current LLMs can handle is limited in areas that require iterative cycles of specialized domain knowledge application, advanced simulations, and data collection through experimentation.

Nevertheless, LLMs have demonstrated a wide range of capabilities, including extensive knowledge and programming skills, which offer a wide range of potential applications. In the field of chemistry, various explorations have been attempted [4, 5], such as extracting some knowledge of compound properties [6, 7] or synthetic routes [8]. However, tasks essential to compound design, such as accurate recognition and manipulation of structural formulae such as SMILES [9], handling of 3D structures, or numerical manipulation of 3D physicochemical properties, are challenging [10]. While there are

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

reports of use cases where collaboration with external toolsets has resulted in drug candidates that slightly differ from existing compounds [11], the acquisition of novel scaffolds for active compounds remains elusive. This reflects an inherent challenge of computational drug discovery, where the true activity of designed compounds remains unknown until they are synthesized and tested in actual assays.

Meanwhile, the automation of programming tasks by LLMs has the potential to revolutionize research and education in chemistry [12, 13]. Individuals with no expertise in medicinal chemistry could initiate computational drug discovery efforts using LLMs, and those with limited programming skills could easily create code to perform compound design calculations using LLMs. Broadening the base of individuals involved in drug discovery science is beneficial to the advancement of the field.

With this in mind, the LLM Drug Discovery Challenge was conceived as a competition to further explore the feasibility of using LLMs in drug discovery. The challenge aimed to harness the collective ingenuity of participants to propose molecular structures of drug candidates for a given drug target using LLM-based workflows. In addition to exploring the practical utility of LLMs in drug discovery, the challenge aimed to identify current limitations and pave the way for future advances in this rapidly evolving field. We recruited participants for this challenge via social media and attracted nine participants, along with a panel of five experts for evaluation. In this paper we describe the design of the challenge, the proposed methodologies, the medicinal chemistry validity of the proposed compound sets and the outlook for the future.

2 Contest Design and Rules

The contest commenced on March 30, 2023, and concluded on June 4, 2023. The details of the rules were posted on our GitHub repository [14], and participants were recruited online, mainly on Twitter.

2.1 General Rules

Participants were required to propose ten candidate compounds predicted to be active for a specified drug target, using processes in their workflow where the LLM either suggests or selects compounds. Each participant was allowed one submission set, with the option to overwrite submissions during the submission period.

2.1.1 Submission Requirements

The submissions were to include:

- Markdown output of interactions with the LLM or a detailed description of the process if not in dialogue format, uploaded to GitHub Gist or similar platforms.
- Code for any computational processing executed outside the LLM, uploaded to GitHub Gist or similar platforms.
- Ten proposed compounds in SMILES format.

2.1.2 Evaluation System

In this event, instead of evaluating activity through actual assay experiments, it was decided to assess whether participants could propose compounds feasible from a medicinal chemistry perspective using LLMs. The evaluation comprised peer reviews among participants and assessments by a panel of judges, combining the score based on the judges' expert evaluation of the molecular structures, and the score assessing the effective utilization of the LLMs (Supplementary Figure 1: the overall scoring formulation). The following five criteria were set as scoring items.

- Chemical stability or Reactivity
- Synthetic accessibility
- Synthetic amenability to diverse derivatizations
- Structural alerts
- Potential for bioactivity

Table 1: The participants' backgrounds and types of methods used by all participants.

Player No.	Backgrounds	Method Type	Ranking
1	Chemoinformatics (MS student)	Hybrid	2nd place
2	Bioinformatics (MS student)	Hybrid	3rd place
3	Computer science (PhD student)	Only LLMs	
4	IT professional	Only LLMs	
5	Orgchem, compchem (PhD student)	Hybrid	1st place
6	Computer Science (Undergrad)	Only LLMs	
7	Compchem (Industry)	Hybrid	Top LLM utilization
8	Orgchem professional	Only LLMs	
9	Orgchem professional	Only LLMs	

A panel of five judges with diverse expertise in medicinal chemistry, computational chemistry, and structural biology was assembled to evaluate the submissions. The evaluation was conducted using a dedicated web application [15] (Supplementary Figure 2) prepared specifically for this scoring task.

2.2 Target Molecule

While our event was held independently from another drug discovery competition, Critical Assessment of Computational Hit-finding Experiments (CACHE) 4th [16, 17], the selection of the drug discovery target followed the lead of CACHE 4th, with the TKB domain of CBL-B protein [18] being chosen. Approximately 900 patented compounds known to bind to this target were provided as references. The goal was to propose novel active compound candidates with low structural similarity to these known compounds, assessed using metrics like the Tanimoto coefficient on Morgan fingerprints. Known protein-ligand complex structure was available (8GCY.PDB [19]), and targeting the same binding site was mandatory (Supplementary Figure 3). If utilizing commercial compound libraries, we recommended the Enamine Hit Locator Library [20] as a default set.

3 Results

The evaluation was carried out based on the scoring criteria outlined in the previous section, leading to the identification of four award winners. The scores and awards for each participant are shown on the contest website [21]. A summary of the methodologies employed by all nine participants is presented in Table 1. Here, we delve into the detailed methodologies of the four award winners.

3.1 Winner with the Highest LLM Utilization Score

Through repeated trial and error, the participant determined that the performance and limitations of ChatGPT and other LLMs at the time of the contest would not allow us to create a drug candidate compound with a simple prompt, so he decided to utilize a step-by-step execution application called AutoGPT [22, 23]. AutoGPT is software that allows LLMs not only to simply respond to a given prompt by specifying a final goal, but also to think step-by-step about the sub-goals necessary to achieve that goal and to accomplish them in a continuous process. AutoGPT also allows users to implement and add functionality in Python or shell scripts, in addition to features that did not exist in ChatGPT at the time of the contest, such as a web search function, for example.

He added the ability to perform similarity searches and generate chemical structures using REINVENT as a chemoinformatics plugin to AutoGPT. In order to emphasize scaffold hopping of chemical structures in the contest, REINVENT [24] and AutoDockVina [25]/DockStream [26] were combined to generate chemical structures based on docking scores. He also confirmed that AutoGPT can perform relatively simple programming, such as implementing a Python script to extract specific tag information from an SDF file, by itself.

Finally, by directing AutoGPT to generate chemical structures with high docking scores based on PDB information, although some human support was required, the top 10 compounds were submitted. Although it is not possible to have current LLMs perform all of their tasks autonomously, this methodology may be one approach to drug discovery by LLMs with human support.

3.2 First-place Winner in the Overall Score

The methodology incorporates a uniquely tuned LLM, integration with external tools, and further includes judgments based on his expertise as an organic chemist. The workflow is as follows:

The known 895 ligands and compounds obtained from ChEMBL [27] were docked against the protein derived from PDB:8GCY. The acquired docking scores were evaluated on a five-point scale and were used to fine-tune the LLM (Open-CALM) [28] using LoRA. The model was later employed to predict the binding affinity of new compounds to the target protein.

A compound generation model (STONED) [29] was used to virtually generate compounds. STONED aimed to create compounds with a balance of similarity and novelty compared to known ligands based on them. Various filters were applied to the generated compounds to ensure they met certain criteria such as synthetic accessibility, molecular weight, and distinctiveness from known ligands.

The fine-tuned LLM was utilized to narrow down compounds from Enamine's library [20] and the group of compounds produced by STONED. In the five-point scale evaluation of the docking score, compounds predicted to have the highest binding affinity were extracted, narrowing down approximately 470,000 compounds to 5,614. In the end, 10 compounds that exhibited docking poses similar to the co-crystal of 8GCY were selected.

This methodology integrates docking, machine learning fine-tuning, virtual compound generation, and selection processes to identify potential compounds for further evaluation in drug discovery. The use of LLM, in conjunction with LoRA for fine-tuning, and STONED for compound generation, showcases a multi-faceted approach to harness computational techniques in the realm of medicinal chemistry.

3.3 Second-place Winner in the Overall Score

This methodology aims to generate novel structures starting from compounds with known structures on GPT-4. Submissions were selected in four steps: molecule generation by GPT-4, conversion to synthesizable compounds, activity evaluation by docking, and selection of desirable compounds.

First, the molecule generation process by GPT-4 iteratively generates SMILES of new compounds that improve the evaluation based on known compounds and their evaluation. Four evaluation filters were used: 2 filters to avoid known scaffolds, "Novelty" based on Tanimoto similarity to known compounds, and "Goodness" based on a MolSkill [30]-inspired score function. Prompts are dynamically generated based on reference SMILES that are randomly sampled from the compound library. Then, the generated compounds are added to the compound library. The iterative process resulted in 145 compounds. Then, 47 compounds were extracted that did not contain any known scaffolds and had original MolSkill [30] scores less than -10. The generated compounds were converted to compounds contained in Enamine's Hit Locator Library to avoid the synthesizability issue. Here, the generated compounds were converted to the nearest compound in the library based on Tanimoto similarity with ECFP4. Next, activity evaluation by docking used template docking in Cresset Flare V7's LeadFinder [31]. The known complex structure (PDB:8GCY) was referenced and the binding pose was visually evaluated, and then the activity was determined by VSScore [31]. Finally, The 10 compounds were selected for submission based on the MaxMin method from 18 compounds that had VS scores less than -8 and satisfied the BRENK [32] and PAINS [33] filters.

The methodology showcased GPT-4's capability to commence from known compounds, avoid certain substructures, and propose novel entities. The compounds exist in Enamine, making experimental validation feasible. Manual efforts were invested in scrutinizing the compounds' properties and stereochemistry. However, it may be possible to generate novel and highly active compounds by considering the activity of GPT-4 in the generation stage.

3.4 Third-place Winner in the Overall Score

The methodology employed MERMAID[34], a vital tool for structured compound discovery for the specified drug target, and divided into Generation, Filtration, and Selection phases. MERMAID, a molecular generative model with structural target information, was central to this approach.

MERMAID served as an in-silico hit-to-lead optimization tool, refining compounds based on affinity via Monte Carlo Tree Search. Autodock Vina[35] conducted docking simulations using initial

molecules from the PDB database. Approximately 10k compounds were generated. These compounds underwent filtration based on criteria like molecular weight, ring structure, and synthetic potential. A machine learning-based retrosynthesis analysis tool further narrowed the selection, reducing the pool to 1k compounds. Lastly, LLM selected final candidates based on similarity to known inhibitors. This methodology combined MERMAID's exploration, computational tools, and LLM's capabilities, offering a comprehensive approach to streamline drug discovery. It leveraged machine learning, computational chemistry, and domain-specific knowledge to expedite promising compound identification.

4 Discussion

Various molecular structures were submitted through the distinct methodologies employed by each participant. As depicted in Supplementary Figure 5, compounds proposed solely using LLMs tended not to escape the trend of high similarity with known compounds. On the other hand, the set of compounds proposed using external toolsets tended to yield molecules with promising prospects from a medicinal chemist's perspective, and lower similarity to known compounds, depending on the performance of individual tools and the expertise of the user.

4.1 Challenges in Establishing Evaluation Criteria

The contest was designed to broadly investigate the applicability of LLMs without specifying how they should be used. This situation necessitated that participants strike a balance between two critical aspects: i) efficient selection of promising compounds, ii) utilization of LLMs in the process.

Some participants prioritized the former, incorporating LLMs only partially and leaving room for alternative methods. Conversely, others emphasized the latter, which sometimes led to the submission of less desirable compounds. This diversity in approach can be attributed to the vague guidelines of the contest. While this ambiguity may have encouraged creative explorations, it likely also led to uncertainty among the participants. This suggests room for improvement in the contest's administration and points toward issues to be addressed in future iterations.

4.2 Utilized LLM Methods

A variety of molecular design methods using LLMs have been reported, ranging from those that directly apply OpenAI's models (e.g., ChatDrug [36]) to those that use LLMs to guide the selection of various tools (e.g., ChemCrow [37]). Multimodal approaches that connect molecules with natural language (e.g., CLAMP [38], KV-PLM [39], MolFM [40]) were also observed. However, most of these models are currently limited to simple molecular transformations and are best suited for auxiliary roles in molecular design. In this contest, direct generation of molecules via models like ChatGPT generally yielded less promising results. On the other hand, molecular evaluations using LLMs sometimes showed comparable performance to existing methods employing traditional molecular representations, even offered unique predictive results, indicating a certain level of utility in this contest.

5 Conclusions

The LLM Drug Discovery Contest served as a valuable platform to evaluate the extent to which LLMs can be pragmatically employed in the drug discovery field. Although we have not yet realized the ideal "one-click" solution by LLMs, the contest has likely offered valuable insights for the community, pointing toward future directions for achieving this goal.

Acknowledgments and Disclosure of Funding

We thank all the contest participants and judges in the author list, as well as Prof. Motoyuki Hattori (Fudan University) and Dr. @taxyach for their invaluable contributions as judges. The service prizes were kindly provided by AgroDesign Studios Co. Ltd. and iSiP, Inc. The prize money awarded to the contest winners was sponsored by SHaLX Inc.

References

[1] Navraj S Nagra, Lieven van der Veken, Erika Stanzl, David Champagne, Alex Devereson, and Matej Macak. The company landscape for artificial intelligence in large-molecule drug discovery. *Nat. Rev. Drug Discov.*, August 2023.

[2] Madura K P Jayatunga, Wen Xie, Ludwig Ruder, Ulrik Schulze, and Christoph Meier. AI in small-molecule drug discovery: a coming wave? *Nat. Rev. Drug Discov.*, 21(3):175–176, March 2022.

[3] Gisbert Schneider. Automating drug discovery. *Nat. Rev. Drug Discov.*, 17(2):97–113, February 2018.

[4] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using GPT-4. *arXiv*, page 2311.07361, November 2023.

[5] Kan Hatakeyama-Sato, Naoki Yamane, Yasuhiko Igarashi, Yuta Nabae, and Teruaki Hayakawa. Prompt engineering of GPT-4 for chemical research: what can/cannot be done? *Science and Technology of Advanced Materials: Methods*, 3(1). 2260300.

[6] Clayton W Kosonocky, Claus O Wilke, Edward M Marcotte, and Andrew D Ellington. Mining patents with large language models demonstrates congruence of functional labels and chemical structures. *arXiv*, page 2309.08765, September 2023.

[7] Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *arXiv*, page 2305.18365, May 2023.

[8] Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, Sam Cox, Wibe A de Jong, Matthew L Evans, Nicolas Gastellu, Jerome Genzling, María Victoria Gil, Ankur K Gupta, Zhi Hong, Alishba Imran, Sabine Kruschwitz, Anne Labarre, Jakub Lála, Tao Liu, Steven Ma, Sauradeep Majumdar, Garrett W Merz, Nicolas Moitessier, Elias Moubarak, Beatriz Mouriño, Brenden Pelkie, Michael Pieler, Mayk Caldas Ramos, Bojana Ranković, Samuel G Rodriques, Jacob N Sanders, Philippe Schwaller, Marcus Schwarting, Jiale Shi, Berend Smit, Ben E Smith, Joren Van Herck, Christoph Völker, Logan Ward, Sean Warren, Benjamin Weiser, Sylvester Zhang, Xiaoqi Zhang, Ghezal Ahmad Zia, Aristana Scourtas, K J Schmidt, Ian Foster, Andrew D White, and Ben Blaiszik. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digit. Discov.*, August 2023.

[9] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, February 1988.

[10] Cayque Monteiro Castro Nascimento and André Silva Pimentel. Do large language models understand chemistry? a conversation with ChatGPT. *J. Chem. Inf. Model.*, 63(6):1649–1655, March 2023.

[11] OpenAI. GPT-4 technical report. *arXiv*, page 2303.08774, March 2023.

[12] Glen M Hocky and Andrew D White. Natural language processing models that automate programming will transform chemistry research and teaching. *Digit. Discov.*, 1(2):79–83, April 2022.

[13] Andrew D White, Glen M Hocky, Heta A Gandhi, Mehrad Ansari, Sam Cox, Geemi P Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, and Willmor J Peña Ccoa. Assessment of chemistry knowledge in large language models that generate code. *Digit. Discov.*, 2(2):368–376, April 2023.

[14] Kusuri Murakumo. LLM_DD_Challenge: Dashboard for LLM drug discovery challenge. `https://github.com/souyakuchan/LLM_DD_Challenge/tree/main`. Accessed: 2023-12-01.

[15] Naruki Yoshikawa, Kentaro Rikimaru, and Kazuki Z Yamamoto. Sanitize it yourself: Web-based molecular sanitization for machine-generated chemical structures. *ChemRxiv*, December 2021.

[16] Suzanne Ackloo, Rima Al-Awar, Rommie E Amaro, Cheryl H Arrowsmith, Hatylas Azevedo, Robert A Batey, Yoshua Bengio, Ulrich A K Betz, Cristian G Bologa, John D Chodera, Wendy D Cornell, Ian Dunham, Gerhard F Ecker, Kristina Edfeldt, Aled M Edwards, Michael K Gilson, Claudia R Gordijo, Gerhard Hessler, Alexander Hillisch, Anders Hogner, John J Irwin, Johanna M Jansen, Daniel Kuhn, Andrew R Leach, Alpha A Lee, Uta Lessel, Maxwell R Morgan, John Moult, Ingo Muegge, Tudor I Oprea, Benjamin G Perry, Patrick Riley, Sophie A L Rousseaux, Kumar Singh Saikatendu, Vijayaratnam Santhakumar, Matthieu Schapira, Cora Scholten, Matthew H Todd, Masoud Vedadi, Andrea Volkamer, and Timothy M Willson. CACHE (critical assessment of computational hit-finding experiments): A public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat. Rev. Chem.*, 6(4):287–295, April 2022.

[17] Finding ligands targeting the TKB domain of CBLB. `https://cache-challenge.org/challenges/finding-ligands-targeting-the-tkb-domain-of-cblb`. Accessed: 2023-10-2.

[18] Serah Kimani, Sumera Perveen, Magdalena Szewezyk, Hong Zeng, Aiping Dong, Fengling Li, Pegah Ghiabi, Yanjun Li, Irene Chau, Cheryl Arrowsmith, Dalia Barsyte-Lovejoy, Vijayaratnam Santhakumar, Masoud Vedadi, and Levon Halabelian. Probing the mechanism of cbl-b inhibition by a small-molecule inhibitor. *bioRxiv*, page 2023.05.05.539612, May 2023.

[19] Rcsb Protein Data Bank. RCSB PDB - 8GCY: Co-crystal structure of CBL-B in complex with N-Aryl isoindolin-1-one inhibitor. `https://www.rcsb.org/structure/8GCY`. Accessed: 2023-10-2.

[20] Hit locator library - enamine. `https://enamine.net/compound-libraries/diversity-libraries/hit-locator-library-460`. Accessed: 2023-10-2.

[21] Kusuri Murakumo. LLM_DD_Challenge: Results and prizes. `https://github.com/souyakuchan/LLM_DD_Challenge/tree/main/result`. Accessed: 2023-12-01.

[22] AutoGPT: An experimental open-source attempt to make GPT-4 fully autonomous. `https://github.com/Significant-Gravitas/AutoGPT`. Accessed: 2023-12-01.

[23] Hui Yang, Sifu Yue, and Yunzhong He. Auto-GPT for online decision making: Benchmarks and additional opinions. *arXiv*, page 2306.02224, June 2023.

[24] Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. REINVENT 2.0: An AI tool for de novo drug design. *J. Chem. Inf. Model.*, 60(12):5918–5922, December 2020.

[25] Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. AutoDock vina 1.2.0: New docking methods, expanded force field, and python bindings. *J. Chem. Inf. Model.*, 61(8):3891–3898, August 2021.

[26] Jeff Guo, Jon Paul Janet, Matthias R Bauer, Eva Nittinger, Kathryn A Giblin, Kostas Papadopoulos, Alexey Voronov, Atanas Patronov, Ola Engkvist, and Christian Margreitter. DockStream: a docking wrapper to enhance de novo molecular design. *J. Cheminform.*, 13(1):89, November 2021.

[27] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucl. Acids. Res.*, 40(Database issue):D1100–7, January 2012.

[28] CyberAgent. Open-calm. `https://huggingface.co/cyberagent/open-calm-7b`, 2023. Accessed: 2023-09-29.

[29] AkshatKumar Nigam, Robert Pollice, Mario Krenn, Gabriel dos Passos Gomes, and Alán Aspuru-Guzik. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies. *Chem. Sci.*, 12:7079–7090, 2021.

[30] Oh-Hyeon Choung, Riccardo Vianello, Marwin Segler, Nikolaus Stiefl, and José Jiménez-Luna. Learning chemical intuition from humans in the loop. *ChemRxiv*, March 2023.

[31] Oleg V. Stroganov, Fedor N. Novikov, Viktor S. Stroylov, Val Kulkov, and Ghermes G. Chilov. Lead finder: An approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening. *J. Chem. Inf. Model.*, 48(12):2371–2385, 2008.

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

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

[34] Daiki Erikawa, Nobuaki Yasuo, and Masakazu Sekijima. Mermaid: an open source automated hit-to-lead method based on deep reinforcement learning. *J. Cheminform.*, 13:94, 2021.

[35] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31:455–461, 2010.

[36] Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. ChatGPT-powered conversational drug editing using retrieval and domain feedback. *arXiv*, page 2305.18090, May 2023.

[37] Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv*, page 2304.05376, April 2023.

[38] Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. *arXiv*, page 2303.03363, March 2023.

[39] Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nat. Commun.*, 13(1):862, February 2022.

[40] Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. MolFM: A multimodal molecular foundation model. *arXiv*, page 2307.09484, June 2023.