PharmGPT: Building a GPT-based Chat Model for Drug Discovery by Unifying Heterogeneous Domain Knowledge

Wenbo Zhang[a,b,1], Shuo Sun[a,1], Yixin Liu[a], Peidong Liu[a], Jin Liu[c], Jiancheng Lv[a], Bowen Ke[c], Sen Song[b], Xianggen Liu[a,*]

[a] *College of Computer Science, Sichuan University, Chengdu, 610065, China*
[b] *Laboratory of Brain and Intelligence and Department of Biomedical Engineering, Tsinghua University, Beijing, 100084, China*
[c] *Laboratory of Anesthesia and Critical Care Medicine, Department of Anesthesiology, Translational Neuroscience Center, West China Hospital, Sichuan University, Chengdu, 610041, China*

Abstract

Large language models (LLMs) demonstrate remarkable capabilities in general-purpose natural language generation, image generation, and multi-domain understanding. However, their performance in specific professional domains, such as drug discovery, has been somewhat lackluster. This limitation can be attributed to the absence of domain-specific corpora. Despite the abundance of pharmaceutical data, it is characterized by high specialization, diverse sources, and varying formats, which pose significant challenges for model training. To bridge this gap, we propose a unified data generation framework that transforms heterogeneous data into a textual corpus with standardized formats. Utilizing the generated data, we develop a powerful chat model with comprehensive pharmaceutical knowledge, named PharmGPT.

*Corresponding author.
Email address: liuxianggen@scu.edu.cn (Xianggen Liu)
[1] These authors contributed equally to this work.

Our experimental results illustrate that PharmGPT outperforms ChatGPT and other LLMs on various tasks, including domain-specific question answering, molecule optimization, and synthesis routes prediction, delivering more accurate and professional responses.
Keywords: Large language model, Drug discovery, Heterogeneous knowledge unification

1. Introduction

The large language models (LLMs) have made a significant impact on the field of natural language processing (NLP) [1, 2]. Recently, instruction-following models such as ChatGPT and GPT-4 show notable success in generating human-like responses, making significant progress toward general artificial intelligence. Trained across a diverse array of textual corpora, these models can adeptly produce contextually relevant and nuanced responses to a broad spectrum of natural language queries and prompts. Although LLMs exhibit generalization across various tasks ranging from text generation [1] to reasoning [3, 4] and programming [5], they continue to face limitations in excelling within professional questions in domain-specific fields. For example, an LLM trained on general web text may not have the specialized knowledge or vocabulary to accurately understand and generate responses for topics related to medicine, law, or engineering.

In particular, LLMs currently have limited coverage of the pharmaceutical domain for drug discovery, where the knowledge is often dispersed across diverse data formats. Generally speaking, there are mainly four types of data formats in the field of drug discovery: 1) Basic information about drugs

is typically described in structured formats, such as DrugBank [6]; 2) The advanced discoveries, evidence, and conclusions related to drug functions are reported in the research papers with natural language. 3) Drug properties measured by wet-lab experiments are often recorded and managed in tabular formats. 4) The synthesis routes of drug molecules are represented in tree-structured formats, indicating the step-wise reaction to iteratively obtain the molecule of interest. In addition, data formats in drug discovery also contain gene sequences, 3D conformations, 2D molecule images, and topological connections between targets and drugs. These diverse data formats deviate from natural language, posing challenges in integrating the heterogeneous knowledge into the training corpus of LLMs.

In this paper, we propose a unified data generation framework that transforms heterogeneous data into a textual corpus with standardized formats. We assume that the general LLMs such as ChatGPT are capable of domain-specific question-answering (QA) if the corresponding knowledge is provided in advance. Therefore, we design tailored prompts for different data formats for ChatGPT to leverage this knowledge and generate corresponding dialogue data in QA format. Specifically, ChatGPT is instructed to produce diverse dialogues based on the given knowledge, including the basic information of the drug functions, the advanced research discoveries in literature, the tabular drug properties, and the synthesis routes (Figure 1). By incorporating the generated dialogue data, we effectively merge the heterogeneous domain knowledge into a unified corpus for LLMs.

Utilizing the generated data, we develop a powerful chat model for drug discovery, coined as **PharmGPT**. PharmGPT adopts the LLaMA [7] frame-

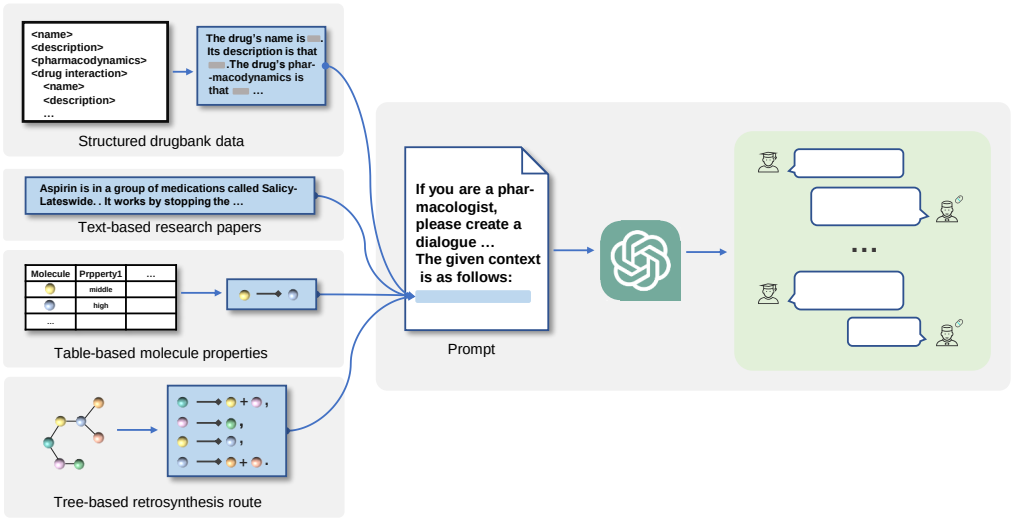

Figure 1: The unified data generation framework that transforms heterogeneous knowledge into a textual corpus with dialogue format.

work and adjusts the vocabulary to better fit the molecule generation tasks. By fine-tuning based on LLaMA, the PharmGPT model empowers users with an intuitive and efficient tool for navigating pharmaceutical knowledge.

To evaluate the performance of PharmGPT, we conduct various QA experiments to compare PharmGPT and other LLMs. In experiments, PharmGPT provides more accurate and insightful responses in domain-specific QA tasks related to drug discovery. LLMs are questioned to accomplish molecule optimization and synthesis route prediction. We observe, except for PharmGPT and ChatGPT, the other LLMs can hardly perform molecule optimization. In addition, ChatGPT fails to predict the synthesis routes of the molecules. By contrast, PharmGPT exhibits reasonable and professional responses for molecule optimization and synthesis routes prediction. These findings demonstrate PharmGPT's potential as an effective computational

tool in drug discovery. In summary, our contributions are as follows:

- We propose PharmGPT, a GPT-based chat model that possesses comprehensive and enriched knowledge in the pharmaceutical domain. To our knowledge, PharmGPT is the *first* computational model that can perform molecule optimization and synthesis route prediction through dialogues.

- We introduce a unified pipeline to build the chat data from heterogeneous knowledge for the training of LLMs. This framework is general and effective to unify heterogeneous domain knowledge, which can benefit not only the pharmaceutical field but also various other domains.

- Experimental results show that PharmGPT offers accurate and professional responses to queries in drug discovery, showcasing its exceptional expertise in domain-specific QA, molecule optimization, and synthesis route prediction. In particular, we conduct two-turn dialogues to show the promising potentials of PharmGPT for molecule optimization, revealing that the multi-turn dialogues stand as a novel way to address the computational problems in drug discovery.

2. Related Work

2.1. Deep learning in drug discovery

Deep learning in drug discovery is a field of artificial intelligence (AI) that leverages deep neural networks to expedite and enhance the process of discovering new drugs and understanding their mechanisms. A large number

of deep learning methods are applied to drug discovery due to their powerful capability of feature extraction and flexibility of model structures [8].

Based on the forms of molecular representation, these methods can be categorized into two main approaches: sequence-based [9] and graph-based methods [10]. Sequence-based methods typically involve the use of convolutional neural networks or recurrent neural networks to process molecular sequences represented as SMILES strings [11]. However, extracting structural information from these sequences can be challenging. The emergence of graph networks provides a new direction for molecular representation. Graph-based methods operate on molecular graphs that consist of atoms and bonds, offering a more comprehensive representation of structural information of molecules. The molecules are usually processed by different graph neural networks, such as graph convolutional network [12] and graph attention network [13, 14], to accomplish the prediction task.

These deep learning models have often been tailored to specific tasks, limiting their versatility. In contrast, this paper aims to create a more general natural language-based dialogue model, enriched with knowledge related to drug discovery. This model enables a more interactive and conversational approach to accessing and retrieving drug-related information and is particularly useful for clinicians and researchers who need to quickly and accurately locate relevant information in real-world scenarios.

2.2. Large language models (LLMs)

Large language models have gained significant attention in recent years due to their impressive capabilities in various natural language processing tasks such as question answering [15, 16], image generation [17, 18] and senti-

ment analysis [19, 20]. One influential work is the introduction of OpenAI's GPT model. The original GPT model demonstrated remarkable language generation abilities by leveraging unsupervised learning on a massive corpus of text data. This work sparked a wave of research and innovation in the field of large language models. Following the success of GPT, subsequent works (GPT2, GPT3 [1]) focused on improving model architectures and training techniques. Additionally, ChatGPT, an extension of GPT, specifically targets conversational applications. It allows for interactive and dynamic conversations with users, providing more engaging and context-aware responses. This work has led to improvements in chatbot systems and dialogue generation.

Afterwards, numerous large models emerged, among which the emergence of the LLaMA [7] model once again caused a considerable sensation. Compared to ChatGPT, the LLaMA model not only has open parameters to the public, but its model parameters are also nearly ten times smaller. At the same time, LLaMA also provides parameter models of different scales (7B - 65B) for tasks with different needs. The emergence of LLaMA provides a more convenient way for the use and research of large language models.

2.3. LLMs for drug discovery

Large-scale language models have exhibited exceptional performance in general domains; however, their performance in specific professional domains such as drug discovery has been relatively lackluster. This deficiency can be attributed to a dearth of domain-specific knowledge. Pharmaceutical knowledge is characterized by high specialization, diverse sources, and varying formats, which contributes to the challenges involved in its collection.

In response to this gap, several endeavors have been made to address this

issue. Med-PaLM [21] is developed by fine-tuning Flan-PaLM using seven medical question-answering datasets. ChatDoctor [22] collected 100k authentic doctor-patient dialogues from the online medical consultation platform HealthCareMagic to fine-tune the LLaMA model. HuaTuo [23] is a Chinese biomedical LLM tuned with curated knowledge from the Chinese medical knowledge graph (CMeKG) [24].

Current models primarily revolve around medical consultation platforms, which is a general and public application in the medical field. We strive to take it a step further by dedicating our efforts to a more specialized and professional pursuit: drug discovery. This specialized large model will offer comprehensive and in-depth insights into drug-related information, empowering users with the knowledge they need to drive advancements and innovations in the pharmaceutical industry.

3. Method

3.1. Heterogeneous domain knowledge in drug discovery

We mainly collect domain knowledge in drug discovery from four heterogeneous data sources to obtain a wide range of information and expertise. These data show diverse formats or structures which are distinct from natural language, and are rarely utilized in the training of previous LLMs. Therefore, integrating these knowledge plays a pivotal role in advancing the understanding of LLMs in the field of drug discovery.

3.1.1. Structured knowledge

Structured knowledge here pertains to the basic information of a drug that is systematically organized in a predefined format. Within the pharmaceutical

realm, structured data plays an integral role in drug discovery and development processes. Here we extract the structured knowledge from the DrugBank [6], renowned for its structured data containing basic drug information about FDA-approved drugs as well as experimental drugs going through the FDA approval process. Although it is one of the world's most widely used reference drug resources, the rich content in DrugBank has not yet been fully utilized in the training of LLMs due to its structured format. Training LLMs on this knowledge can equip them with a comprehensive understanding of the basic information about drugs.

3.1.2. Textual knowledge

Textual data plays a crucial role in NLP applications. It serves as the basis for training language models and building dialogue systems. In the field of drug discovery, textual data, such as research papers and patents, is a rich source of information for drug development. For textual knowledge, we diligently collect recent research papers in the pharmaceutical field. These papers are currently underused in LLMs, but provide a wealth of detailed and specialized knowledge on various aspects of pharmaceuticals, including drug discovery, mechanisms of action, pharmacokinetics, therapeutic applications, and clinical studies. By training LLMs on such data, the models can acquire a comprehensive understanding of the domain-specific terminology, concepts, and research findings. In addition, research articles are continuously published, reflecting the latest advancements and discoveries in the drug field. Training LLMs on these up-to-date papers allows them to capture the most current knowledge and keep pace with the rapidly evolving landscape of pharmaceutical research.

3.1.3. Tabular knowledge

Tabular knowledge encompasses a consistent set of properties applied across different drug molecules. Tabular data is used extensively in medical trials to record and manage data related to drug development and formulation It facilitates the efficient organization, comparison and interpretation of experimental data. We construct a drug-like properties knowledge base of molecules derived from the ZINC [25] and ChEMBL [26, 27] datasets, which contain comprehensive information on various drug properties. Within this tabular knowledge base, we have compiled pertinent drug-like attributes, such as molecular weight, octanol/water partition coefficient (LogP value), topological polar surface area, number of hydrogen bond acceptors, and more. This integration has empowered LLMs to assimilate diverse perspectives and insights into drug properties, thereby enabling them to be capable of drug property optimization.

3.1.4. Tree-structured knowledge

Tree-structured data is valuable in the pharmaceutical domain for representing hierarchical relationships and complex structures. It plays a crucial role in synthesis routes, enabling the representation, analysis, and exploration of chemical reactions and synthesis pathways. We delve into tree-structured knowledge, specifically synthesis routes, which is a critical aspect in drug development that aims to find efficient synthetic pathways of a target molecule by recursively transforming it into easier precursors. The synthesis routes are collected from the USPTO dataset [28, 29], which are structured as tree-like format and have received limited exploration in LLM training. By incorporating these routes into LLM training, the models gain insights into the diverse

range of chemical transformations and synthetic strategies employed in drug synthesis.

3.2. Dialogue dataset generation

To effectively utilize the aforementioned knowledge for fine-tuning LLM, a crucial initial step is the unification of heterogeneous domain knowledge. In order to accommodate diverse data formats, we employ distinct approaches to transform them into linear natural language representations. We formulate tailored prompts to facilitate the generation of dialogue data through ChatGPT. Specifically, we assume the roles of a pharmacologist and a Ph.D. student, employing ChatGPT to generate multi-turn dialogues centered around the given data. The provided data is considered factual, and the answers provided by the pharmacologist during the conversation are expected to be impartial, accurate, and evidence-based. The statistics of generated dialogues are shown in Table 1. The detailed transformation processes corresponding to diverse data formats are elucidated in the following.

3.2.1. Generation based on structured knowledge

For structured knowledge, we extract essential items such as drug names, descriptions, and pharmacological actions from DrugBank. Subsequently, we design a template to embed the extracted items and assemble them into natural language expressions. For example, we transform the structural information of Lepirudin into the following text: "The drug's name is Lepirudin. And its description is that Lepirudin is identical to natural hirudin except for substitution of leucine for isoleucine at the N-terminal end of the molecule and the absence of a sulfate group on the tyrosine at position 63. It is produced

Table 1: Statistics of generated data

Source	Number of generated dialogue	Number of dialogue rounds
Structured knowledge	14594	77064
Textual knowledge	32804	170097
Tabular knowledge	12474	72047
Tree-structure knowledge	15000	81925
Total	74872	401133

via yeast cells." The text is incorporated into the prompts which are then presented into ChatGPT to generate multi-turn dialogues revolved around the provided knowledge.

3.2.2. Generation based on textual knowledge

For textual knowledge, we collect the recent research papers and extract the introduction sections which spans diverse facets like drug development, drug interactions and mechanism of action. Since they are already in the textual form, we seamlessly integrate them into the prompts, which are then used to engage ChatGPT in generating multi-turn dialogues.

3.2.3. Generatinon based on tabular knowledge

We have constructed a tabular knowledge base of molecules from the ZINC and ChEMBL datasets, encompassing a comprehensive array of drug-like attributes. To convert tabular knowledge into language description, we transform it into molecular optimization processes, where the goal is to optimize a given starting molecule towards desirable properties. Specifically, we match the molecules in pairs and calculate their similarity degree. If the similarity exceeds a predefined threshold, we consider the two molecules to be

transformable. Subsequently, we generate a series of molecule transformations for modifying a specific property (e.g., increasing the number of hydrogen bond acceptors in a molecule), thereby achieving modifications in molecular properties. Consider a scenario where the similarity between molecules A and B is substantial, and A possesses more hydrogen bond acceptors than B. In this case, we regard this as an optimization process for increasing hydrogen bond acceptors, expressed as follows: "There are some instances that can increase the number of hydrogen bond acceptors. A $\rightarrow$ B." In this way, we effectively connect individual molecular properties to adjustments in molecule properties. These transformations are grouped into the designed prompts to generate corresponding dialogue data as shown in Figure 2.

3.2.4. Generation based on tree-structured knowledge

The synthesis routes of drug molecules are organized in a hierarchical tree-like structure, providing universally applicable information for biomedical research. These pathways originate from validated synthetic routes, showcasing chemical transformations and reaction sequences. We reconstructed each layer of the tree nodes into linear transformation equations, resulting in the disassembly of the tree structure into a sequence of equations. This approach allows us to simulate the synthesis process by representing it as a sequence of transformations. For example, for molecule A, its synthetic route can be expressed as follows: "B + C $\rightarrow$ A, D + E $\rightarrow$ C.". These transformations are incorporated into the prompts to generate corresponding dialogue data.

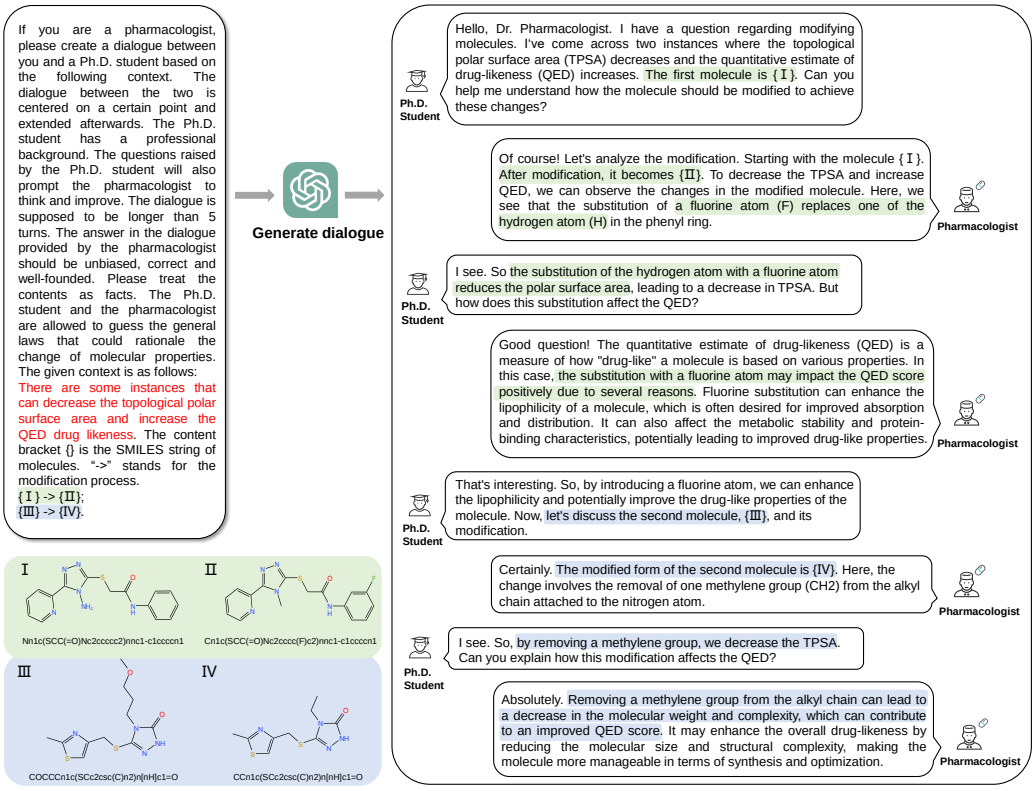

Figure 2: Illustration of dialogue generation based on tabular knowledge

3.3. Training of PharmGPT model

Our model is developed by fine-tuning the LLaMA-7B model with the LoRA plugin [30]. LoRA is a plugin that helps large models adapt to datasets. The advantage is that it is very fast during finetune, and the resulting model is also very small, about 30M. Our purpose of fine-tuning LLaMA is to train the model parameters of the LoRA plugin.

We merge all the generated dialogue data from different sources to fine-tune PharmGPT model. The model is trained on a Nvidia 3090 GPU with 24 GB RAM for 3 epochs. The hyperparameters employed in the process of

training are as follows: a learning rate of 3×10^{-4}, a batch size of 128, a maximum sequence length of 256 tokens, a lora rank of 8, a lora alpha of 16 and a lora dropout of 0.05.

4. Experiment

4.1. Baseline models

LLaMA. LLaMA is a large open-source language model by Meta AI. Due to its exceptional performance achieved with minimal resource utilization, LLaMA has garnered extensive recognition and popularity within the community. LLaMA offers various versions, including 7B, 13B, 33B, and 65B, based on the parameter sizes. For our study, we utilize LLaMA-7B as the foundational model for PharmGPT.

Alpaca. Alpaca is an instruction-following fine-tuning model built upon LLaMA and developed by Stanford. In comparison to proprietary models from OpenAI, Alpaca demonstrates comparable performance while being more readily accessible.

Vicuna. Vicuna is an open-source chatbot developed through the fine-tuning of LLaMA using user-shared conversations gathered from ShareGPT. Initial assessment, employing GPT-4 as the evaluator, demonstrates that Vicuna-13B attains a quality level exceeding 90% compared to ChatGPT and Google Bard. Furthermore, Vicuna-13B consistently surpasses other models, including LLaMA and Stanford Alpaca, in over 90% of instances.

ChatGPT. ChatGPT is an advanced language model developed by OpenAI. It is specifically designed for generating human-like text in a conversational manner. With its ability to understand and respond to prompts,

ChatGPT enables interactive and engaging conversations with users.

4.2. Evaluation metrics

We conduct a comprehensive analysis to compare PharmGPT with baseline models, including domain-specific answer generation, expert evaluation, drug property prediction and synthesis route prediction of drug molecules.

- **Domain-specific question answering**. To assess the quality of question answering (QA) in the pharmaceutical field, we construct a collection of QA dialogues via ChatGPT. As we utilize genuine knowledge data as prompts for ChatGPT, which boasts substantial text generation capabilities, we hold the view that the dialogue data produced by ChatGPT stands as a reasonably dependable representation of our ground truth. Therefore, we use previous prompts from DrugBank and research papers (described in 3.2.1 and 3.2.2) to regenerate new dialogues. In these dialogues, the student's query is used as the question and the pharmacologist's answer is regarded as the ground truth. Then we feed these questions into PharmGPT and baseline models, and compare the responses provided by these models to the ground truth answers. BLEU [31] and BertScore are adopted as metrics to evaluate model performance.

- **Expert evaluation**. We also invited experts from the pharmaceutical domain to assess the responses of different models. For the medical QA tasks, we adopt usability and smoothness metrics proposed by Wang [23], where usability reflects the medical expertise of a specific response and smoothness represents the ability as a language model. Each aspect

is scored from 1 to 5, representing extreme low, low, neutral, high, and extreme high, respectively. Note that the evaluation is conducted in a blind fashion, where the experts do not know the identity of the model.

- **Molecule optimization**. Given a molecule, ask the model to optimize it for desired properties (e.g., rotatable bonds (Rotbonds), H-bond acceptors (HBA) and topological polar surface area (TPSA), octanol/water partition coefficient (LogP value), the fraction of C atoms that are SP3 hybridized (C-SP3)). Then we test whether the relevant properties of the modified molecule in the response meet the requirements and obtain the accuracy of molecule optimization.

- **Synthesis route prediction of molecules**. To test the ability of the model for synthesis route prediction, we ask the model to provide the synthetic routes of target molecules. Then we compare them with the reference route to judge the correctness of the answer, and get the accuracy of synthesis route prediction.

4.3. Results

We first evaluate the capacity of domain-specific question answering and Table 2 presents the performance of different models. The primary metric for comparison is the BLEU score, wherein PharmGPT demonstrates superior performance in the QA task utilizing both knowledge sources (i.e. DrugBank and research articles). In addition, we calculate the F1 score based on the BERTScore and observe that PharmGPT exhibits modest enhancement compared to the baseline models. It is worth noting that BERTScore relies on BERT's representations, which may not possess specialized knowledge in the

pharmaceutical domain. As a result, the improvement in this metric may not be as pronounced. Overall, the convergence of the two metrics shows superior capacity of PharmGPT in question answering in pharmaceutical domain.

Table 2: Performance of domain-specific answer generation

Source	Model	BLEU	BERTScore		
			Recall	Precision	F1 scores
DrugBank	LLaMA	3.18	0.85	0.86	0.85
	Alpaca	9.93	0.88	0.89	0.89
	Vicuna	7.50	0.89	0.87	0.88
	ChatGPT	6.48	**0.90**	0.86	0.88
	PharmGPT	**12.88**	0.89	**0.90**	**0.89**
Paper	LLaMA	5.60	0.85	0.86	0.85
	Alpaca	9.82	0.88	**0.88**	0.88
	Vicuna	7.32	0.88	0.86	0.87
	ChatGPT	5.62	**0.89**	0.85	0.87
	PharmGPT	**11.80**	0.88	0.88	**0.88**

Table 3: Results of expert evaluation for different models

Model	Usability	Smoothness
LLaMA	2.09	2.95
Alpaca	3.35	3.87
Vicuna	3.79	4.14
ChatGPT	4.08	4.36
PharmGPT	**4.34**	**4.53**

For expert evaluation, we recruit 2 annotators with pharmaceutical background to assess the usability and smoothness of the responses from different models. We sample 100 questions and score the corresponding answer from

Table 4: Accurary of drug property modification and synthesis route prediction. "-" indicates that the model is not applicable to the task.

Dialogue	Model	Property				Synthesis
		QED & C-SP3	Rotbonds & LogP	QED & RotBonds	C-SP3 & LogP	
Single-turn	LLaMA	-	-	-	-	-
	Alpaca	-	-	-	-	-
	Vicuna	-	-	-	-	-
	ChatGPT	0.29	0.13	0.22	0.43	-
	PharmGPT	**0.44**	**0.23**	**0.38**	**0.52**	**0.16**
Two-turn	LLaMA	-	-	-	-	-
	Alpaca	-	-	-	-	-
	Vicuna	-	-	-	-	-
	ChatGPT	0.46	0.18	0.37	0.60	-
	PharmGPT	**0.50**	**0.24**	**0.41**	**0.62**	-

1 to 5. The average scores are shown in Table 3. PharmGPT shows higher usability than other models, reflecting better pharmaceutical expertise. As a language generation model, the responses of PharmGPT are also smoother than others, which is consistent with the automatic metrics in Table 2.

Additionally, we conduct experiments to probe the ability of molecule optimization. For each property being tested, we sample 100 molecular SMILES strings which are not presented in the training set and ask the model to modify these molecules to change the desired properties. Notably, we choose property combinations absent in the training set as the test objectives. Once we obtain the modified SMILES strings, we verify whether the resulting molecules have indeed altered the property compared to the original molecules. Table 4 presents the accuracy of different models in modifying properties. The term "single-turn accurac" pertains to the accuracy of the model's answer within a single interaction. In addition, we have introduced a novel metric

known as "*two-turn accuracy*". This innovative approach involves informing the model about the correctness of its initial response and prompting it to generate a new response if the initial one is incorrect. The two-turn accuracy is then calculated based on the responses obtained in these two interaction rounds.

Remarkably, PharmGPT's single-turn accuracy has substantially exceeded that of the baseline models by an impressive margin of over 10% in four property combinations, demonstrating its superiority in molecular optimization. Moreover, the two-turn accuracy has been further improved, indicating the efficacy of the proposed metric. PharmGPT's two-turn accuracy still remains superior to that of the baseline model, albeit with a comparatively lesser improvement in the accuracy of the second round compared to ChatGPT. This might be attributed to ChatGPT's robust interactive capabilities, an aspect we intend to further develop in the future.

Furthermore, we investigate PharmGPT's potential in the synthesis routes prediction. We sample 100 molecular SMILES strings that do not exist within the training set, and ask the model to provide the corresponding synthetic routes for each molecule. As shown in Table 4, PharmGPT successfully predicts the synthesis routes for 16 molecules out of 100, while other baseline models face significant challenges and are not applicable to this task.

4.4. Case study

4.4.1. Molecule optimization by PharmGPT

We showcase a case on modification of molecular properties in Figure 3. Given a molecule, the user queries the models to design a new molecule that increases the LogP value and the fraction of C-SP3. In response, ChatGPT

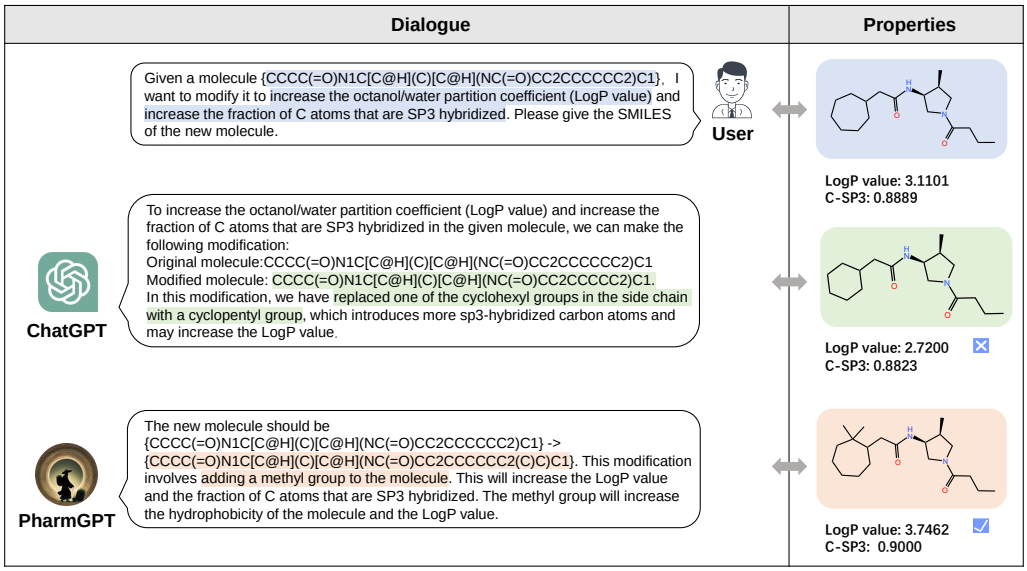

Figure 3: Case study for molecule optimization.

provides inaccurate molecule with lower LogP value and fraction of C-SP3. On the contrary, PharmGPT produces a new molecule that fulfills the specified criteria. Notably, PharmGPT not only produces the correct molecule but also provides a coherent and logical explanation for the modification process, suggesting that the addition of methyl groups will effectively elevate the LogP value and enhance the fraction of C-SP3.

4.4.2. Synthesis route prediction by PharmGPT

We present an instance of the synthesis route prediction in Figure 4. As observable, the initial step underscores PharmGPT's proficiency in identifying the SMILES strings of the target molecule. Subsequently, PharmGPT furnishes two simpler precursor molecules for the synthesis of the target molecule. Given that one of the intermediate molecules remains challenging to obtain di-

rectly, PharmGPT expounds the synthesis route of this intermediate molecule and further transforms it into a more accessible precursor. We also test the other LLMs such as ChatGPT and Vicuna, but they all fail to accomplish the synthesis route prediction task.

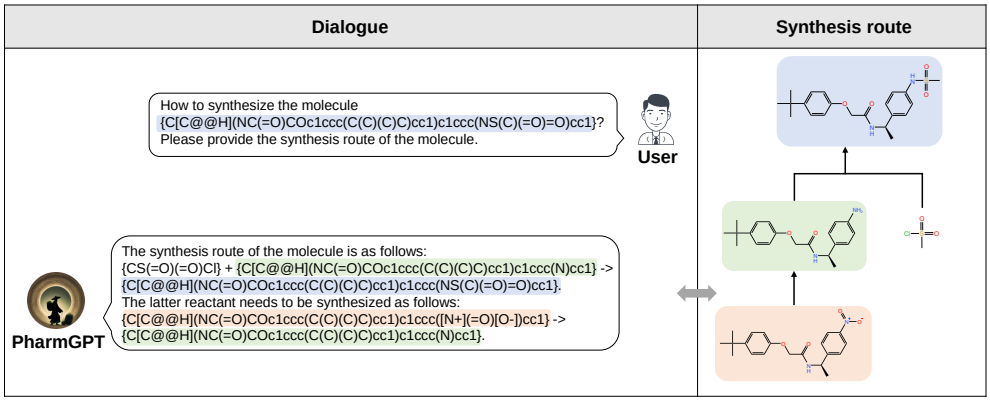

Figure 4: Case study for synthesis route prediction.

5. Conclusion and discussion

In this paper, we develop the PharmGPT model, a chat model enriched with a wide range of heterogeneous domain knowledge of drug discovery. We present a unified data generation framework that involves creating tailored prompts and and then generating corresponding dialogue data through ChatGPT. This framework is general and effective to transforms diverse data formats into a standardized textual corpus, which is useful in many fields including but not limited to pharmaceutical domain. By fine-tuning PharmGPT on the unified data, PharmGPT shows superior pharmaceutical expertise. Experimental results demonstrate that PharmGPT outperforms

other baseline models on various tasks in the field of drug discovery. Notablely, PharmGPT is a pioneering computational model capable of engaging in molecule optimization and synthesis route prediction through dialogue interactions. Additionally, we introduce a novel evaluation method known as "two-turn accuracy" to validate the promising potential of PharmGPT in molecule optimization. This approach shows the potential of multi-turn dialogues as an innovative approach to addressing computational challenges in drug discovery.

Despite the advantages of our method, it has limitations. First, the practical utility of the model comes with the inherent risk of generating incorrect responses. Second, although our model surpasses many baseline models such as ChatGPT, its accuracy, particularly in synthetic route prediction, may not be consistently remarkable. This indicates the necessity for further exploration and refinement in subsequent endeavors. Third, the interpretability of PharmGPT is limited because of the opaque nature of deep learning. In the future, we will incorporate neural symbolism [32] in PharmGPT to improve its interpretability.

6. Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants (82273784 and 62206192), the Natural Science Foundation of Sichuan Province under Grant (2023NS-FSC1408), the 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYYC21002), the Clinical Research Innovation Project, West China Hospital, Sichuan University (2019HXCX06), the National Natural Science Foundation of China

under Grants 61836004.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 1877–1901.

[2] D. W. Otter, J. R. Medina, J. K. Kalita, A survey of the usages of deep learning for natural language processing, IEEE Transactions on Neural Networks and Learning Systems 32 (2) (2021) 604–624.

[3] K. Shuang, J. Guo, Z. Wang, Comprehensive-perception dynamic reasoning for visual question answering, Pattern Recognition 131 (2022) 108878.

[4] J. Qin, Z. Yang, J. Chen, X. Liang, L. Lin, Template-based contrastive distillation pretraining for math word problem solving, IEEE Transactions on Neural Networks and Learning Systems (2023).

[5] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., Evaluating large language models trained on code, arXiv preprint arXiv:2107.03374 (2021).

[6] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., Drugbank 5.0: a

major update to the drugbank database for 2018, Nucleic Acids Research 46 (D1) (2018) D1074–D1082.

[7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[8] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The rise of deep learning in drug discovery, Drug Discovery Today 23 (6) (2018) 1241–1250.

[9] A. Lavecchia, Deep learning in drug discovery: opportunities, challenges and future prospects, Drug Discovery Today 24 (10) (2019) 2017–2032.

[10] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems 32 (1) (2021) 4–24.

[11] D. Weininger, Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules, Journal of Chemical Information and Computer Sciences 28 (1) (1988) 31–36.

[12] Q. Zhang, J. Chang, G. Meng, S. Xu, S. Xiang, C. Pan, Learning graph structure via graph convolutional networks, Pattern Recognition 95 (2019) 308–318.

[13] X.-b. Ye, Q. Guan, W. Luo, L. Fang, Z.-R. Lai, J. Wang, Molecular substructure graph attention network for molecular property identification in drug discovery, Pattern Recognition 128 (2022) 108659.

[14] Q. Lv, G. Chen, Z. Yang, W. Zhong, C. Y.-C. Chen, Meta learning with graph attention networks for low-data drug discovery, IEEE Transactions on Neural Networks and Learning Systems (2023).

[15] F. Xu, Q. Lin, J. Liu, L. Zhang, T. Zhao, Q. Chai, Y. Pan, Y. Huang, Q. Wang, Moca: Incorporating domain pretraining and cross attention for textbook question answering, Pattern Recognition 140 (2023) 109588.

[16] C. Chen, D. Han, C.-C. Chang, Caan: Context-aware attention network for visual question answering, Pattern Recognition 132 (2022) 108980.

[17] Z. Tan, X. Yang, Z. Ye, Q. Wang, Y. Yan, A. Nguyen, K. Huang, Semantic similarity distance: Towards better text-image consistency metric in text-to-image generation, Pattern Recognition (2023) 109883.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[19] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognition 136 (2023) 109259.

[20] Y. Zhou, L. Liao, Y. Gao, R. Wang, H. Huang, Topicbert: A topic-enhanced neural language model fine-tuned for sentiment classification, IEEE Transactions on Neural Networks and Learning Systems 34 (1) (2023) 380–393.

[21] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, Nature (2023) 1–9.

[22] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, Y. Zhang, Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, Cureus 15 (6) (2023).

[23] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, T. Liu, Huatuo: Tuning llama model with chinese medical knowledge, arXiv preprint arXiv:2304.06975 (2023).

[24] B. Odmaa, Y. Yunfei, S. Zhifang, D. Damai, C. Baobao, L. I. Sujian, Z. Hongying, Preliminary study on the construction of chinese medical knowledge graph, Journal of Chinese Information Processing 10 (2019) 1–9.

[25] T. Sterling, J. J. Irwin, Zinc 15–ligand discovery for everyone, Journal of Chemical Information and Modeling 55 (11) (2015) 2324–2337.

[26] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, et al., Chembl: towards direct deposition of bioassay data, Nucleic Acids Research 47 (D1) (2019) D930–D940.

[27] M. Davies, M. Nowotka, G. Papadatos, N. Dedman, A. Gaulton, F. Atkinson, L. Bellis, J. P. Overington, Chembl web services: streamlining access to drug discovery data and utilities, Nucleic Acids Research 43 (W1) (2015) W612–W620.

[28] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain, Chemical Science 11 (1) (2020) 154–168.

[29] S. Genheden, E. Bjerrum, Paroutes: towards a framework for benchmarking retrosynthesis route predictions, Digital Discovery 1 (4) (2022) 527–539.

[30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.

[31] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[32] X. Liu, Z. Lu, L. Mou, Weakly supervised reasoning by neuro-symbolic approaches, Compendium of Neurosymbolic Artificial Intelligence (30) (2023).