ChatGPT in Drug Discovery

Gaurav Sharma[*a], Abhishek Thakur[b]

[a] Department of Chemistry, Michigan State University, USA

[b] Center for Biophysics and Computational Biology, Temple University, USA

**Corresponding Author:* Gaurav Sharma

gaurav.sharmapsit@gmail.com

Abstract: ChatGPT is a language model developed by OpenAI. It is a machine learning model that has been trained on a large dataset of human language, allowing it to generate human-like text. It can be used for a variety of natural language processing tasks such as language translation, text summarization, and question answering. In the current work we have discussed the application of ChatGPT in drug discovery.

Keyword: ChatGPT, computational chemistry, drug discovery.

Introduction:

Chat GPT is a state-of-the-art language model that is developed by OpenAI.[1] It is a deep neural network that has been trained on a large corpus of text, which enables it to generate human-like responses to a wide range of prompts.[1] Chat GPT utilizes the transformer architecture, which allows it to handle long-term dependencies and context in its language generation. It can be fine-tuned to specific tasks such as language translation, question answering and text completion, and can generate human-like text for various applications.[1, 2] Chat GPT is a notable example of the advancement in natural language processing and machine learning. Since, it has been trained by up till 2021, it processes the user questions and try to give us good information back on that topic.[1] Hence the information provided by Chat GPT is not novel it will just provide the information uptill 2021.[1, 2]

Drug discovery is the process of identifying and developing new medications to treat disease.[3] It is a complex, multi-disciplinary field that involves various steps such as target identification, lead discovery, preclinical development, clinical trials, and regulatory approval. The process typically starts with the identification of a biological target that is involved in a disease process.[3] Researchers then use various techniques such as high-throughput screening, computational chemistry, and structure-based drug design to identify small molecules or biologics that can interact with the target and modulate its activity.[3-5] These lead compounds are then subjected to preclinical testing in various animal models and cell lines to evaluate their efficacy and safety before moving on to clinical trials in humans.[6, 7] The goal of drug discovery is to identify safe and effective treatments for a wide range of diseases.[3, 6] Computational based drug discovery process if shown in Figure 1.[3]

Drug discovery using computational chemistry is the application of computer-based methods to assist in the discovery of new medications.[5, 7, 8] It involves the use of computer simulations and modeling techniques to predict the properties and interactions of potential drug

molecules. These methods can be used to understand the structure and behavior of proteins and other biological targets, and to design and optimize new compounds that can bind to these targets and modulate their activity.[9-11] Computational chemistry plays a vital role in modern drug discovery by providing a cost-effective and efficient way to screen large numbers of compounds, predict their potential efficacy and toxicity, and identify new lead compounds for further development.[3, 12-14] It also enables researchers to design new drugs that are more selective and have fewer side effects, as well as to understand the mechanisms of drug action at the molecular level, and to optimize the pharmacokinetics and pharmacodynamics of drugs.

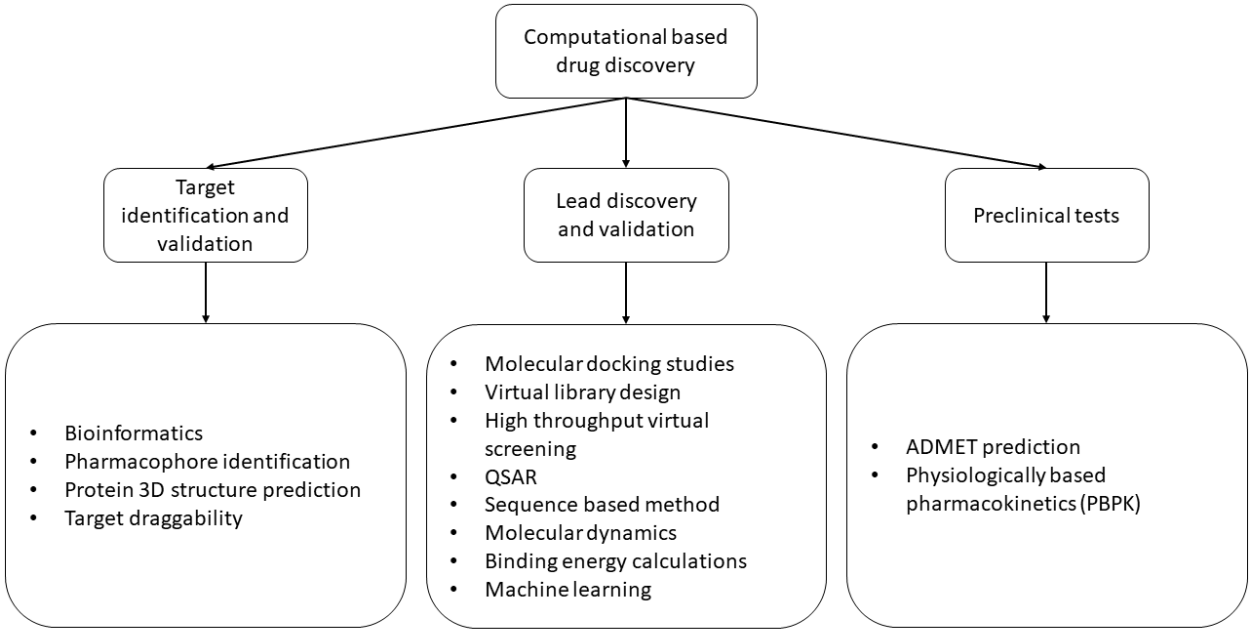

Figure 1: Computational based drug discovery process.

Drug discovery is a very difficult and complex process for several reasons: (1) Biological complexity: The human body is an incredibly complex system, and understanding how it works at the molecular level is difficult. Diseases are caused by a variety of factors, and finding a single drug that can treat a disease effectively is challenging;[15] (2) High failure rate: The majority of drugs that enter clinical trials fail to be approved. This is because they are found to be ineffective, or they have unacceptable side effects during testing; (3) Time-consuming and expensive: The drug discovery process is long and costly. It can take up to 15 years and cost billions of dollars to bring a new drug to market;[16] (4) Stringent regulatory requirements: The process of drug development is heavily regulated, and drugs must go through a series of preclinical and clinical trials before they can be approved for use; (5) Unpredictable nature of drug development: Even after a drug has been identified and developed, it's not certain that it will work as expected in human trials, sometimes it can have unexpected side effects; (6) Lack of new targets: Many of the known drug targets have already been exploited, and finding new targets is becoming increasingly difficult.[3, 15] Despite these challenges, researchers continue to work towards developing new and effective drugs. Advances in technology and computational chemistry, as well as a better

understanding of disease biology, are helping to make the drug discovery process more efficient and cost-effective. In the current work, we have explored the usage and limitation of Chat GPT in drug discoveries.

Results and discussions: To study the Chat GPT we have devised computational questions which can be applied in computational drug discovery process.

1. **Compute the compound multiplicity:** when the multiplicity of Zn was asked Chat GPT gave the correct answer. The answer of Chat GPT is as follows: “Zinc is a transition metal with the atomic number 30, it has 30 electrons in its electronic configuration. The electronic configuration of Zn is [Ar] 3d^10 4s^2. Since all the electrons are paired in the orbitals, the multiplicity of Zn is 1 (singlet). This means that all the electrons in the HOMO are paired and the molecule has no unpaired electrons.” Therefore, Chat GPT can be used to determine multiplicity of various compounds in quantum mechanics calculations using density functional theory (DFT).

2. **Convert smiles to files:** Chat GPT successfully determine the smiles of Tylenol (Acetaminophen), hence it can be used to covert one file format to another. The Chat GPT answer was “The chemical name of Tylenol is Acetaminophen and its molecular formula is C8H9NO2, its Simplified molecular-input line-entry system (SMILES) notation for Tylenol is "CC(=O)NC1=CC=C(C=C1)O". This notation represents the atoms and their connections in the molecule.”
3. **Generation of input files for gaussian and other software:** Chat GPT successfully produces the input file for gaussian software. The input file is shown in Figure 2.

```
# HF/6-31G(d)

Acetaminophen optimization

0 1
C  -0.81617   -0.56170    0.00000
C  -1.54282    0.68142    0.00000
C  -0.12103    1.26552    0.00000
C   1.20888    0.26552    0.00000
C   0.48223   -1.06170    0.00000
C  -1.95342   -1.06170    0.00000
O   2.31523    0.26552    0.00000
N   0.12103   -2.06170    0.00000
O  -2.65777   -1.06170    0.00000
H  -2.38111    0.58142    0.00000
H  -2.38111   -1.64170    0.00000
H   0.12103    2.26552    0.00000
H   1.95342    0.58142    0.00000
```

Figure 2: Gaussian software input file of Tylenol generated by Chat GPT

4. **Input file for AutoDock docking:** ChatGPT successfully gave the input file for AutoDock software and also explained the parameters mentioned in it (Figure 2).

Copy code

```
receptor = protein.pdbqt
ligand = ligand.pdbqt
center_x = 15.0
center_y = 15.0
center_z = 15.0
size_x = 30.0
size_y = 30.0
size_z = 30.0
exhaustiveness = 8
num_modes = 9
```

Figure 3: Input file of autodoc software generated by Chat GPT.

5. **Finding the pdb files:** It can also be useful in PDB search. For instance, it can provide the PDB IDs of metalloprotein, protein bound with DNA, spike protein, etc. However, Chat GPT fails to answer complex questions. For example,

   Question: PDB ID list of Zinc tranporters
   Chat GPT: I'm sorry, but I do not have a specific list of PDB IDs for zinc transporters. The Protein Data Bank (PDB) is a large database that contains structural information for a wide variety of proteins, including transporters. You can use the PDB website to search for specific proteins or families of proteins, including zinc transporters, by keyword or other criteria. You can also use PDB ID to access the structure of a protein.

   Question: PDB ID of Zinc ion channels
   Chat GPT: I do not have a specific list of PDB IDs for zinc ion channels, however you can use the PDB website to search for specific structures by keyword or other criteria. For instance, you can search for "Zinc ion channel" and it will give you a list of structures related to that search. Additionally, some examples of PDB IDs for known zinc ion channels include:
   3K4W: Zinc-bound structure of a bacterial voltage-gated ion channel
   4J9F: Crystal structure of a putative zinc-sensitive channel
   5F5Z: Structure of the zinc-bound conformation of a bacterial voltage-gated ion channel
   6B5N: Structure of a voltage-gated zinc-sensitive ion channel
   6EZY: Structure of the zinc-bound conformation of a bacterial voltage-gated ion channel
   Keep in mind that the PDB is constantly being updated, so new structures may be added over time.

6. **Can give FASTA sequence but cannot compare the FASTA sequence:** In this case ChatGPT was unable to suggest the FASTA sequence. However, to suggests that the fasta sequence can be obtained from the UniProt database.

7. **ADMET properties:** in this case it can only give the properties of the known drugs which can easily obtained from various drugs databases.

8. **Help in literature search:** A major application of ChatGPT can be in literature search and listing the specific findings related to topic.

9. **Code for docking:** interestingly, it can help to write basic code which can be used (with slight modification) extensively in many aspects of drug discovery.

10. **Check for plagiarism:** it can be used to detect plagiarism which makes the process of research and publishing the research work smoothly.

11. **Machine learning and data analysis:** it can be integrated into a program or application using its API, which allows developers to input text and receive a response generated by

the model. The API can be accessed using programming languages such as Python, JavaScript, and C#. Additionally, the OpenAI team has also released a pre-trained version of the model, which can be fine-tuned for specific tasks using a process called transfer learning.

12. **Write a poem about computational chemistry:** A fun way to conclude that ChatGPT can write a computational chemistry related poem as well.

13. **Innovation:** innovation is a major part of the research field which can be done only by asking different question. In this aspect ChatGTP can be very useful to researchers in proposing novel ideas.

Future prospects of ChatGPT in drug discovery:

1. **Identifying and validating new drug targets:** ChatGPT can be fine-tuned on a dataset of scientific literature and used to generate summaries of the latest research on a given disease or biological target. This can help researchers quickly identify new potential targets or gain a better understanding of the current state of research in a specific area.

2. **Designing new drugs:** ChatGPT can be fine-tuned on a dataset of known drug-like molecules and used to generate new chemical structures with similar properties. This can help researchers identify new lead compounds that have a higher chance of success in pre-clinical and clinical studies.

3. **Optimizing drug properties:** ChatGPT can be used to predict the pharmacokinetics and pharmacodynamics of new drugs, and to support the virtual screening of chemical libraries in early-stage drug discovery.

4. **Assessing toxicity:** ChatGPT can be fine-tuned on a dataset of toxicity data and used to predict the potential toxic effects of new drugs.

5. **Generating drug-related reports and papers:** ChatGPT can be fine-tuned on a dataset of drug-related papers and used to generate reports and papers that summarize the current state of research in a specific area. ChatGPT can be fine-tuned on a dataset of known drug-like molecules and used to generate new chemical structures with similar properties. This can help researchers identify new lead compounds that have a higher chance of success in pre-clinical and clinical studies.

Although, it has various advantages it cannot perform extensive scientific calculations like computing RMSD, RMSF, PCA, clustering etc. For these complexes work a trained and experience human involvement is required. It's important to note that ChatGPT is just one tool among many that are used in drug discovery, and it is not a substitute for experimental validation and clinical trials. However, by providing a cost-effective and efficient way to process large amounts of data and generate new knowledge, ChatGPT can assist researchers in making more

informed decisions and accelerate the drug discovery process. It's important to note that ChatGPT is just one tool among many that are used in drug discovery, and it is not a substitute for experimental validation and clinical trials. However, it can significantly speed up and improve the drug development process by providing a cost-effective and efficient way to process large amounts of data and generate new knowledge.

Disadvantage:

While ChatGPT can be a powerful tool in drug discovery, there are also some potential disadvantages to using this technology:

1. Reliance on the quality and availability of data: ChatGPT is only as good as the data it has been trained on. If the data is incomplete, biased, or inaccurate, the model's predictions may not be reliable.
2. Lack of experimental validation: ChatGPT can generate predictions and hypotheses, but it cannot perform experiments or measure the properties of compounds. Therefore, the predictions made by the model need to be validated experimentally.
3. Limited understanding of the underlying biology: While ChatGPT can generate human-like text, it does not understand the underlying biology of the systems it is simulating. Therefore, the predictions made by the model may not always reflect the true complexity of the systems.
4. Limited interpretability: ChatGPT, like other machine learning models, can be difficult to interpret, and it's not always clear how the model arrived at a particular prediction.
5. Limitations in handling uncertainty: ChatGPT is a deterministic model, it cannot account for the uncertainty in the data and predictions.
6. Lack of transparency: ChatGPT is a black box model, it is difficult to understand and explain the internal workings of the model, which can make it difficult to trust the model's predictions.

It's important to keep these limitations in mind when using ChatGPT for drug discovery and to validate the predictions made by the model experimentally, and to use it in combination with other tools and techniques to overcome these limitations.

Conclusion:

In conclusion, ChatGPT is a powerful language model that can assist in the field of drug discovery. By processing and generating human-like text, it can help researchers quickly identify new potential targets, gain a better understanding of the current state of research, design new drugs, and optimize the pharmacokinetics and pharmacodynamics of new drugs. The model can be fine-tuned on specific datasets and used to generate new knowledge that can support decision-making in early-stage drug discovery. However, it's important to note that ChatGPT is just one tool among many that are used in drug discovery, and it is not a substitute for experimental validation and clinical trials. In addition, it failed to perform complex computational calculations like simulating and analyzing the molecular level studies. A major question is any of this is ethical? Or is it asking someone to write your thesis? The type or amount of work ChatGPT can do right now is equal to All the questions I asked ChatGPT is either enhancing what I have already done, or it is giving

you starting point where you still need to do more to complete the research. The work done by ChatGPT in this paper in unethical to do and the answers are already available on the internet. However, I would say that ChatGPT make my life easier by giving a starting point. I strongly believe as technology gets better, ChatGPT will be a valuable tool in helping us becoming more efficient in our research. Nevertheless, the use of ChatGPT in drug discovery is a promising area of research, as it has the potential to significantly speed up and improve the drug development process.

References:

1. Aydın, Ö.; Karaarslan, E. J. A. a. S., OpenAI ChatGPT generated literature review: Digital twin in healthcare. **2022**.
2. Jiao, W.; Wang, W.; Huang, J.-t.; Wang, X.; Tu, Z. J. a. p. a., Is ChatGPT a good translator? A preliminary study. **2023**.
3. Ou-Yang, S.-s.; Lu, J.-y.; Kong, X.-q.; Liang, Z.-j.; Luo, C.; Jiang, H. J. A. P. S., Computational drug discovery. **2012,** *33* (9), 1131-1140.
4. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. J. P. r., Computational methods in drug discovery. **2014,** *66* (1), 334-395.
5. Gangadevi, S.; Badavath, V. N.; Thakur, A.; Yin, N.; De Jonghe, S.; Acevedo, O.; Jochmans, D.; Leyssen, P.; Wang, K.; Neyts, J. J. T. j. o. p. c. l., Kobophenol A inhibits binding of host ACE2 receptor with spike RBD domain of SARS-CoV-2, a lead compound for blocking COVID-19. **2021,** *12* (7), 1793-1802.
6. Leelananda, S. P.; Lindert, S. J. B. j. o. o. c., Computational methods in drug discovery. **2016,** *12* (1), 2694-2718.
7. Madhyastha, H.; Madhyastha, R.; Thakur, A.; Kentaro, S.; Dev, A.; Singh, S.; Kumar, H.; Acevedo, O.; Nakajima, Y.; Daima, H. K. J. C.; Biointerfaces, S. B., c-Phycocyanin primed silver nano conjugates: Studies on red blood cell stress resilience mechanism. **2020,** *194*, 111211.
8. Sharma, G.; Song, L. F.; Merz, K. M. J. J. o. C. I.; Modeling, Effect of an Inhibitor on the ACE2-Receptor-Binding Domain of SARS-CoV-2. **2022**.
9. Thakur, A.; Sharma, G.; Badavath, V. N.; Jayaprakash, V.; Merz Jr, K. M.; Blum, G.; Acevedo, O. J. T. j. o. p. c. l., Primer for designing main protease (Mpro) inhibitors of SARS-CoV-2. **2022,** *13* (25), 5776-5786.
10. Thakur, A.; Dan, N.; Bhakat, S.; Jayaprakash, V.; Banerjee, S. J. A.-C. A. i. M. C., 5'Adenosine Monophosphate-Activated Protein Kinase Modulators as Anticancer Agents. **2016,** *16* (8), 961-972.
11. Sharma, G.; Merz, K. M. J. J. o. c. t.; computation, Mechanism of Zinc Transport through the Zinc Transporter YiiP. **2022,** *18* (4), 2556-2568.
12. Siddique, M. U. M.; Thakur, A.; Shilkar, D.; Yasmin, S.; Halakova, D.; Kovacikova, L.; Prnova, M. S.; Stefek, M.; Acevedo, O.; Dasararaju, G. J. E. j. o. m. c., Non-carboxylic acid inhibitors of aldose reductase based on N-substituted thiazolidinedione derivatives. **2021,** *223*, 113630.
13. Thakur, A.; Timiri, A. K. J. W. J. o. P. S., Designing of potential new aromatase inhibitor for estrogen dependent diseases: A computational approach. **2014**, 13-24.
14. Thakur, A.; Mandal, S. C.; Banerjee, S. J. J. o. A.; Studies, M., Compounds of natural origin and acupuncture for the treatment of diseases caused by estrogen deficiency. **2016,** *9* (3), 109-117.

15. Kumar, N.; Hendriks, B. S.; Janes, K. A.; de Graaf, D.; Lauffenburger, D. A. J. D. d. t., Applying computational modeling to drug discovery and development. **2006,** *11* (17-18), 806-811.
16. Macalino, S. J. Y.; Gosu, V.; Hong, S.; Choi, S. J. A. o. p. r., Role of computer-aided drug design in modern drug discovery. **2015,** *38*, 1686-1701.