Evaluation of search-enabled Pre-trained Large Language Models on retrieval tasks for the PubChem Database

Ash Sze and Soha Hassoun
Tufts University

October 18, 2024

Supporting File 1

Detailed Results on the Protocol

To provide additional details for each protocol, we describe our experiences with GPT-4o in more detail, highlighting values and limitations that are specific to each case. Please note that Protocol #3 is detailed as examples in the main manuscript.

Protocol 1. The first protocol identifies genes and proteins that interact with losartan. Such data in PubChem originates from several sources including DrugBank (Wishart *et al.*, 2008), Drug-Gene Interaction database (Freshour *et al.*, 2021), and others. The PubChem web interface guides the user to go to section 16.2 (Chemical-Target Interaction), and filter against DrugBank interactions. The gold results (see full results in appendix) include several Cytochrome enzymes, UDP-glucuronosyltransferases, and others.

Based on prompting, GPT-4o provides correct yet incomplete results. The response is missing certain entries, e.g., Cytochrome P450 2C19, Cytochrome P450 2C8, and ATO- dependent translocase ABCB1. GPT-4o elaborates on each interaction with information external to PubChem and DrugBank such as Wikipedia (see partial results in Figure S1). The names of some interactions and proteins differ slightly from PubChem, indicating the results were acquired from external sources or training data. While the PubChem protocol had a strict definition of what qualifies as an interaction, and where the interactions must be sourced from, GPT-4o interpreted interactions as broader, and therefore included information about downstream effects of these interactions, listing "RAAS" and "Plasma Renin Activity" in the list of proteins and genes affected. Enhanced prompting for this protocol involved excluding downstream effects and protein or gene information and emphasizing the need for data to only be from PubChem. Unfortunately, any engineering to enhance the prompt

led to incorrect answers with non-PubChem sources which strayed further from the established gold answer.

There is currently no method to retrieve the same information through programmatic access. Therefore, the silver answer is not available.

Protocol 2. The second protocol aims to find drug-like compounds similar to losartan based on a two-dimensional (2-D) similarity search using PubChem. The PubChem web interface guides users to perform a "Similar Structures Search" and apply Lipinski's rule of 5 filters. The gold answer includes a detailed list of compounds that match the specified criteria (see full results in the appendix).

GPT-4o attempts to provide a list of compounds that are structurally similar to losartan while adhering to Lipinski's rule of five. However, the results are incorrect when compared to the gold answer. None of GPT-4o's outputs, such as Candesartan and Irbesartan, match the outputs from the gold answer. Looking into the external links provided by GPT-4o show that the model interpreted losartan as all types of losartan, including losartan Sodium. Enhanced prompting for this protocol involved specifying losartan as only "losartan" and not other variations of the drug. This enhanced prompt only shortened the output, leaving it with Candesartan, Irbesartan, and Valsartan, which were still incorrect compared to the gold standard.

To retrieve similar information programmatically, we used PUG to perform a 2-D substructure search for losartan with CID 3961 and applied filters for Lipinski's rule of 5. The programmatic access involved using a URL to perform the search and obtain a list key, which was then used to request a detailed list of hit CIDs. This method provided comprehensive and accurate results that align with PubChem's standards, but upon comparison with the gold answer, the retrieved information filtered on broader criteria, leading to a longer list of hit targets inclusive of the gold answer. This silver answer highlights a significant difference between searches done on PUG and the PubChem web interface.

We also prompted GPT-4o with a programmatic access approach to provide a PUG link for the 2-D similarity search. The prompt included specific criteria, such as molecular weight, hydrogen bond donors and acceptors, and partition coefficient limits according to Lipinski's rule of five. GPT-4o generated the silver answer PUG URL. Although this is a multi-step protocol, requiring the user to manually input the generated list key for the final result, the programmatic prompt effectively provided the most relevant PUG URL templates for each step for detailed data access (refer to Figure S2 for a full protocol summary).

Protocol 4. The fourth protocol aims to get bioactivity data for hit compounds identified through a substructure search using a specific SMILES string in PubChem (Figure S2). The data originates from various sources within PubChem's database. The PubChem web interface guides users to draw and search using the SMILES string "C1=CC=C(C=C1)C2=CC=CC=C2C3=N[N]N=N3" on the PubChem Sketcher, search under the "substructures" tab, click "Bioactivities" under "Linked datasets," and download the linked data to view the bioactivity results (see full results in the appendix).

GPT-4o is limited in answering multimodal questions, such as those involv-

ing SMILES formulas and drawn structures. When GPT-4o is given the SMILES formula, it cannot directly perform a structure search on PubChem. Instead, it instructs users on how to conduct the search and provides limited information through external links. As a result, GPT-4o can only provide a summary of the expected key information types, including AID, Activity Outcomes, Activity Concentrations, and Activity Names, but not the direct bioactivity data itself. Attempting to enhance the prompt led to code generation which created hypothetical data unrelated to actual PubChem data.

To retrieve similar information programmatically, we utilized PUG to perform a substructure search using the given SMILES string. This was followed by performing a search for bioactivities for each CID from the initial search results. Alternatively, a CSV file of all bioactivities of CIDs can be downloaded if all CIDs are listed in the same link. This method involved using specific URLs to first obtain a list of CIDs from the substructure search and then retrieve detailed bioactivity data for each CID. This approach is labor-intensive and more difficult to interpret than the web interface, but provides highly comprehensive results that align with PubChem.

We also prompted GPT-4o with a programmatic access approach to provide a PUG link for the substructure search using the given SMILES string. The prompt included specific criteria to return CIDs, load the list key, and access assay summaries for the given CIDs. GPT-4o generated the correct PUG URLs, allowing for accurate retrieval of compound and bioactivity data, corresponding to the silver answer (refer to Figure S2 for a full protocol summary).

Protocol 5. The fifth protocol aims to find drugs that target the human type-1 angiotensin II receptor gene. The data originates from several sources within PubChem, including DrugBank. The PubChem web interface guides users to search for "type 1 angiotensin II receptor," select the "Genes" tab and click the entry with "Human," navigate to the 7.1 "Chemical-Gene Interactions" subsection, sort by "Data Source" to view all interacting compounds from DrugBank, and download the necessary sections (see full results in the appendix).

GPT-4o provides a list of drugs that interact with the gene encoding the human type-1 angiotensin II receptor. The response is correct but incomplete. GPT-4o lists the drugs Valsartan, Olmesartan, Telmisartan, Irbesartan, Candesartan, and Azilsartan medoxomil, which is only 7 out of the 14 total drug interactions from DrugBank. Attempting to enhance the prompt by asking for a complete list, or all of the data lead to faulty code generation or an output unchanged from the initial output.

There is no programmatic way to directly access the gene-chemical interactions table. There is no silver response.

Protocol 6. The sixth protocol aims to get the bioactivity data of all chemicals tested against the human type-1 angiotensin II receptor (AT1R) and its rat ortholog. The data originates from various sources within PubChem. The PubChem web interface guides users to access the protein summary page for "type 1 angiotensin II receptor," navigate to the "tested compounds" section, download the list of tested compounds and their bioactivity data, and repeat

the process for "P29089 (Norway rat)" under the orthologous protein section (see full results in the appendix).

GPT-4o provides a summary of the bioactivity data for the most important chemicals tested against the human type-1 angiotensin II receptor and its rat ortholog. GPT-4o cannot directly search PubChem and thus only provides possible bioactivity types, a general overview of the AT1R receptor, and listings of human AT1R interactions for compounds losartan, Olmesartan, and Candesartan, and rat AT1R interactions for compounds Valsartan and Irbesartan. GPT-4o does not explain how it shortlisted these compounds, and descriptions of the compounds are generic and do not reflect the gold answer. Enhanced prompting by re-querying and asking for exact lists of compounds did not improve the quality of the responses for this protocol.

To retrieve similar information programmatically, we utilized PUG to perform a concise bioactivities search using the gene IDs 185 and 81638, corresponding to AGTR1 and AGTR1b respectively. This method involved finding the relevant gene IDs and using URLs to obtain detailed bioactivity data.

We also prompted GPT-4o with a programmatic access approach to provide PUG links for concise gene searches on gene ID 185 and gene ID 81638. The prompt included specific criteria to return a summary of bioactivity data for these gene IDs. GPT-4o generated incorrect PUG URLs due to an invalid input domain (refer to figure S5 for a full protocol summary).

Protocol 7. The seventh protocol aims to find compounds annotated with specific classifications or ontological terms using PubChem. The data originates from PubChem's chemical database. The PubChem web interface guides users to the Classification Browser, selecting "MeSH" for classification and "Compound" for data, searching for "Antihypertensive agents" and saving as "1," repeating for "anti-arrhythmia agents" and saving as "2," and using the AND operator to search for results matching both classifications (see full results in the appendix).

GPT-4o cannot perform a direct search on PubChem and is thus unable to use the Boolean search function required for this protocol. The original protocol requires the user to create two separate searches and use the Boolean operator AND to find results that match both statements. While PubChem has 68 results for this search, none of them correspond to GPT-4o's answers. GPT-4o instead generates hypothetical datasets of chemicals and interactions, which do not accurately reflect the actual data available in PubChem. Enhanced prompting did not improve the quality of the responses for this protocol.

There is currently no way to programmatically access classifications or search by MeSH annotations via PUG-REST. The PUG-REST tutorial acknowledges this as a feature that may be added in the future based on demand. Thus, there is no silver answer for this protocol through programmatic access.

Protocol 8. The eighth protocol aims to find stereoisomers and isotopomers of a compound using an identity search in PubChem. The data originates from PubChem's comprehensive chemical database. The PubChem web interface guides users to search for "CID 60846 structure," select "Identity" and select "Same Isotope" in "Settings" for stereoisomers and "Same stereo" for stereoiso-

topomers and download the results as a CSV file (see full results in the appendix).

GPT-4o provides a list of stereoisomers and isotopomers for Valsartan (CID 60846). The response includes the (S)-enantiomer as a commonly known stereoisomer, and Valsartan-d3 and [3H]Valsartan as isotopomers. However, the names and formatting of these chemical compounds differ from PubChem's data, indicating discrepancies. The response does not match the gold answer from PubChem. Attempting to enhance the prompt by changing "find" to "list" or "retrieve" led to faulty code generation. Re-querying for the correct answer leads to a change in output compounds that was still incorrect.

To retrieve similar information programmatically, we used PUG to perform a fast identity structure search for stereoisomers and isotopomers using the CID 60846. This method involved setting "identity_type" as "same_stereo" for stereoisotopomers and "same_isotope" for stereoisomers and using specific URLs to obtain the silver answer which matched the gold answer.

We prompted GPT-4o with a programmatic access approach to provide PUG links for a fast identity search using the same stereo and same isotope settings for CID 60846. However, the output generated by GPT-4o was incorrect, resulting in PUG-REST faults due to invalid operations. (refer to figure S7 for a full protocol summary)

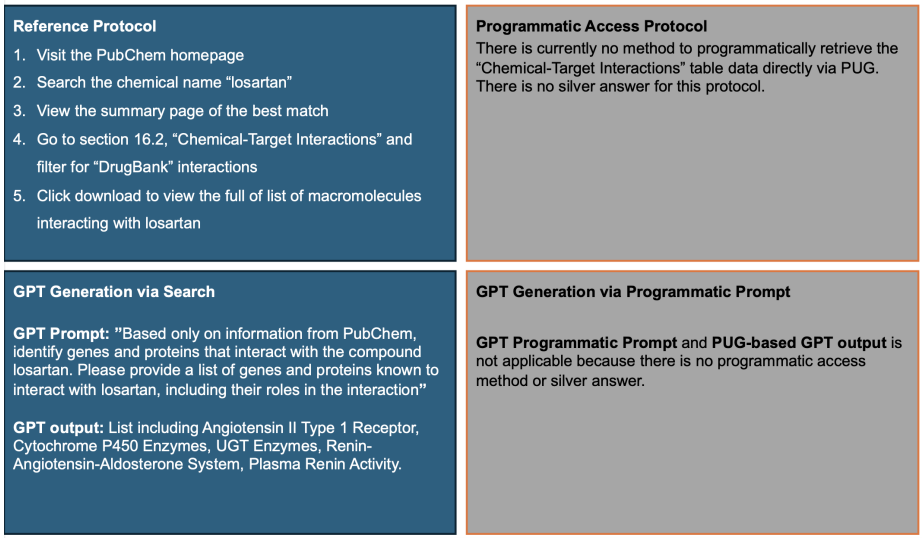

Figure 1: Protocol #1 for finding genes and proteins that interact with a given compound, made specific for losartan.

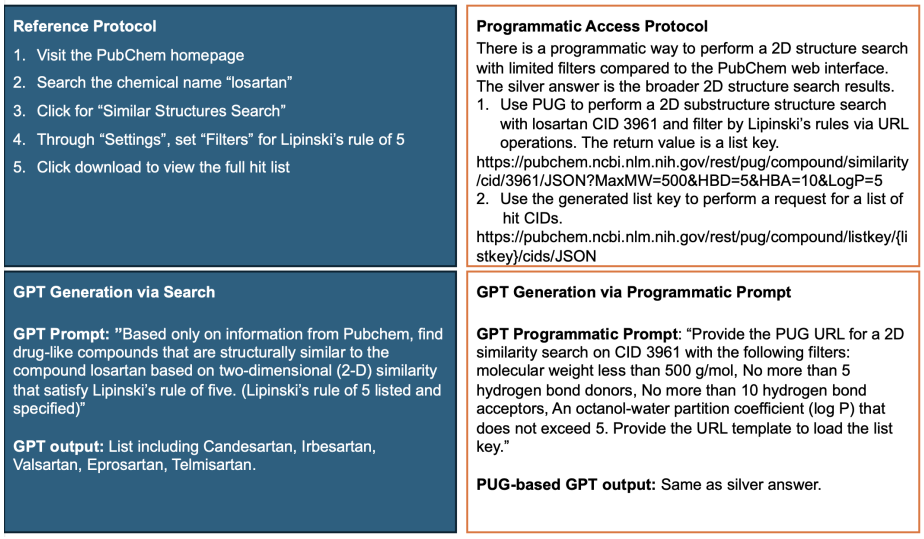

Figure 2: Protocol #2 for finding drug-like compounds similar to a query compound through 2-D similarity search, made specific for losartan filtered through Lipinski's rule of 5.

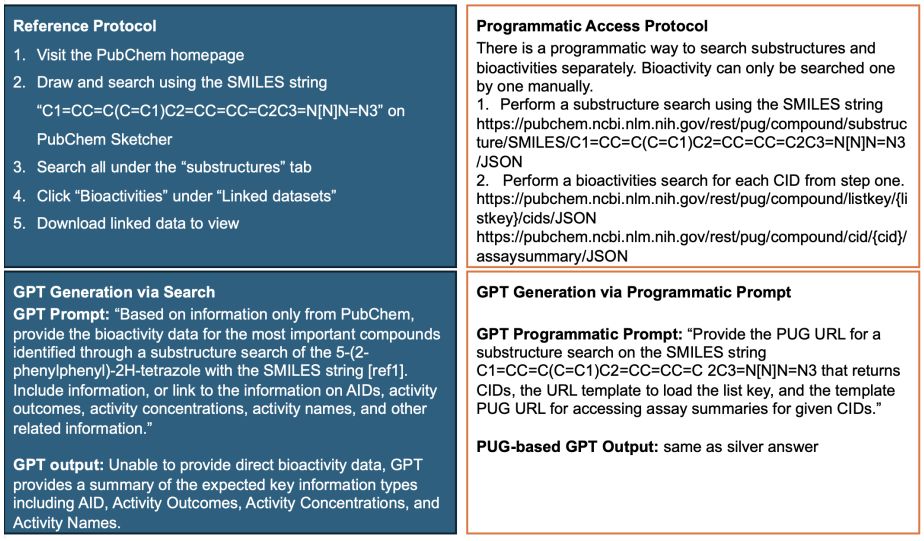

Figure 3: Protocol #4 for getting the bioactivity data for the hit compounds from substructure search, made specific for C1=CC=C(C=C1)C2=CC=CC=C2C3=N[N]N=N3.

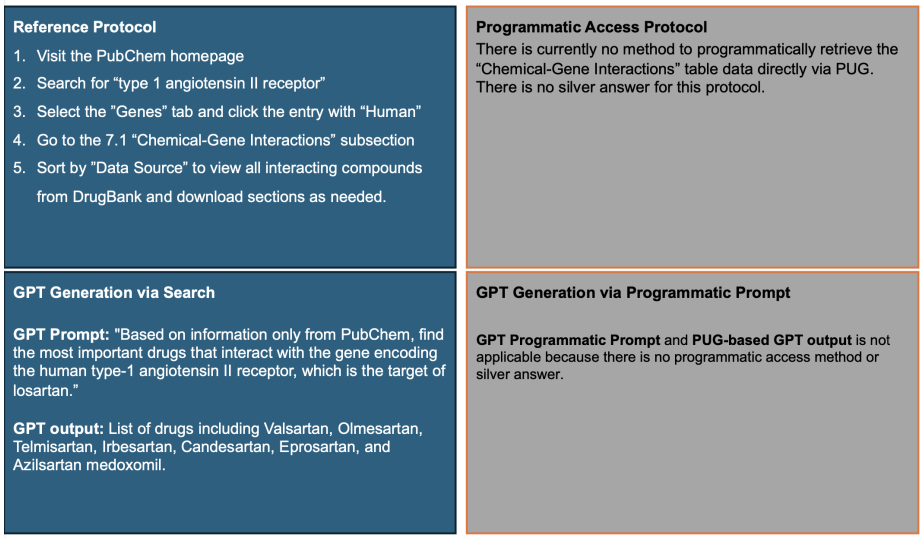

Figure 4: Protocol #5 for finding drugs that target a particular gene, made specific for the gene type 1 angiotensin receptor (human).

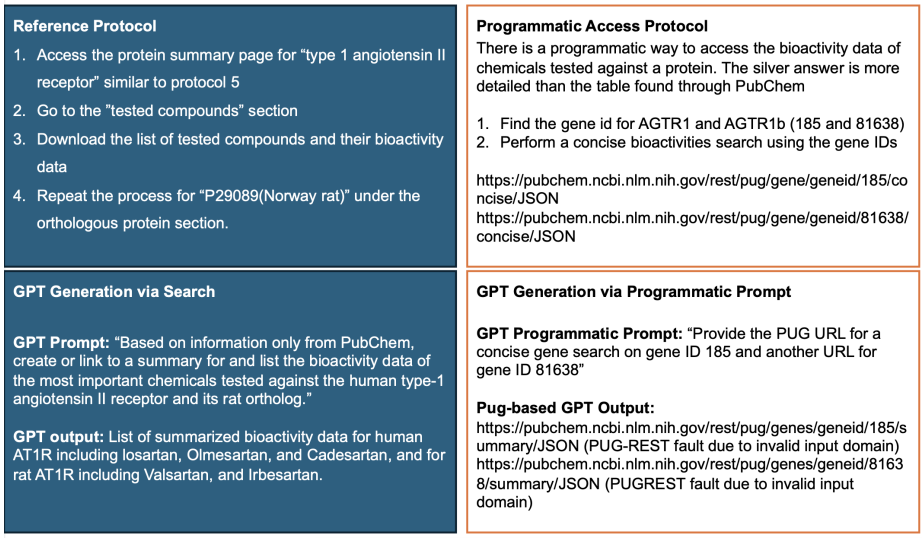

Figure 5: Protocol #6 for getting bioactivity data of all chemicals tested against a protein, made specific for the protein type 1 angiotensin II receptor (human and Norway rat).

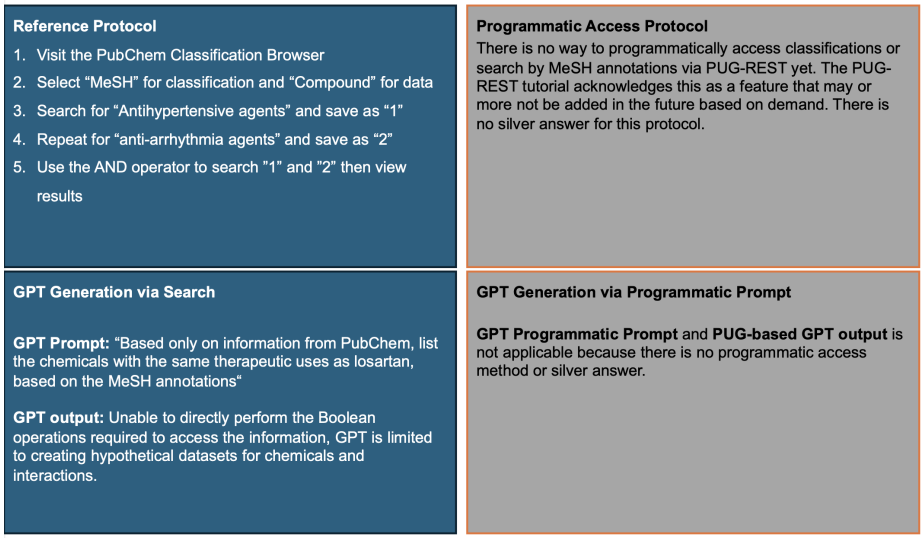

Figure 6: Protocol #7 for finding compounds annotated with classifications or ontological terms, made specific for "antihypertensive agents" and "anti-arrhythmia agents."

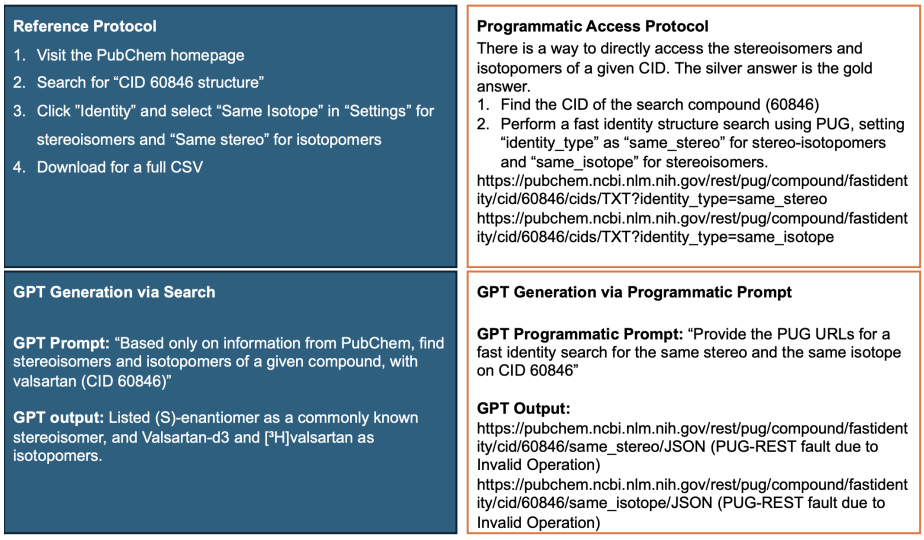

Figure 7: Protocol #8 for getting stereoisomers and isotopomers of a compound through identity search, made specific for a search on "CID 60846 structure."

References

Freshour, S. L., Kiwala, S., Cotto, K. C., Coffman, A. C., McMichael, J. F., Song, J. J., Griffith, M., Griffith, O. L., and Wagner, A. H. (2021). Integration of the drug–gene interaction database (dgidb 4.0) with open crowdsource efforts. *Nucleic acids research*, **49**(D1), D1144–D1151.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, **36**(suppl$_1$), $D901 - -D906$.