# Target Discovery Agent with Large Language Model

David Narganes-Carlon
Senior Data Scientist
Ph.D. Biotechnology and Machine Learning

# What is an LLM-based Agent? The idea

Article

https://doi.org/10.1038/s42256-024-00832-8

## Augmenting large language models with chemistry tools

Received: 13 September 2023

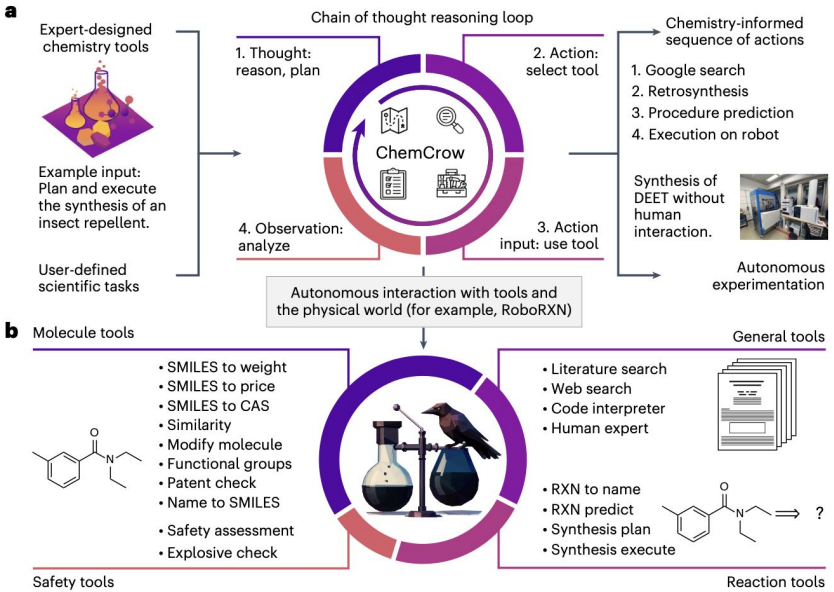
Accepted: 27 March 2024

Published online: 8 May 2024

Check for updates

Andres M. Bran[1,2,6], Sam Cox[3,4,6], Oliver Schilter[1,2,5], Carlo Baldassari[5], Andrew D. White[3,4] & Philippe Schwaller[1,2]

Large language models (LLMs) have shown strong performance in tasks across domains but struggle with chemistry-related problems. These models also lack access to external knowledge sources, limiting their usefulness in scientific applications. We introduce ChemCrow, an LLM chemistry agent designed to accomplish tasks across organic synthesis, drug discovery and materials design. By integrating 18 expert-designed tools and using GPT-4 as the LLM, ChemCrow augments the LLM performance in chemistry, and new capabilities emerge. Our agent autonomously planned and executed the syntheses of an insect repellent and three organocatalysts and guided the discovery of a novel chromophore. Our evaluation, including both LLM and expert assessments, demonstrates ChemCrow's effectiveness in automating a diverse set of chemical tasks. Our work not only aids expert chemists and lowers barriers for non-experts but also fosters scientific advancement by bridging the gap between experimental and computational chemistry.

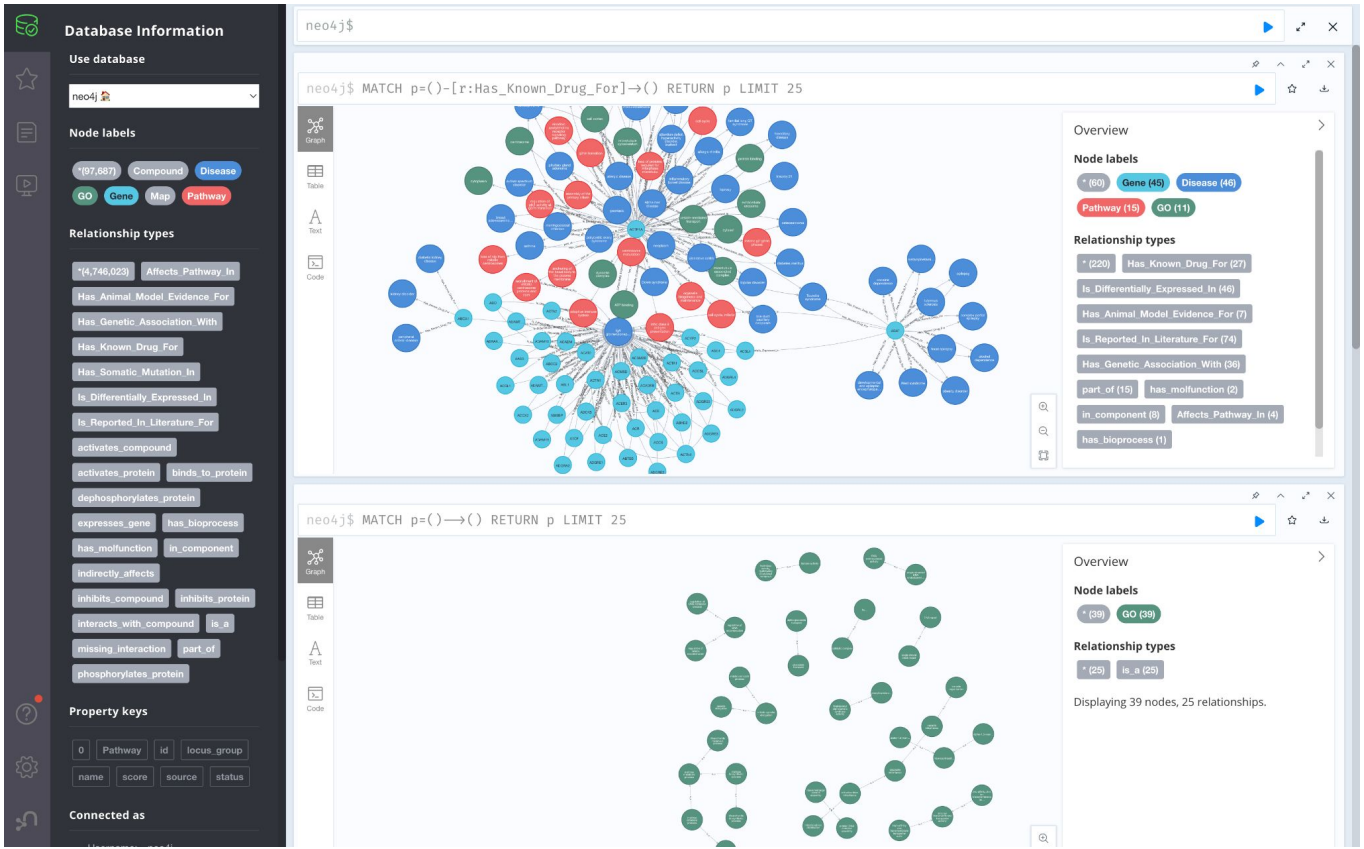



https://arxiv.org/pdf/2304.05376

# My solution: Chat GPT with 4 tools ChemCrow-like

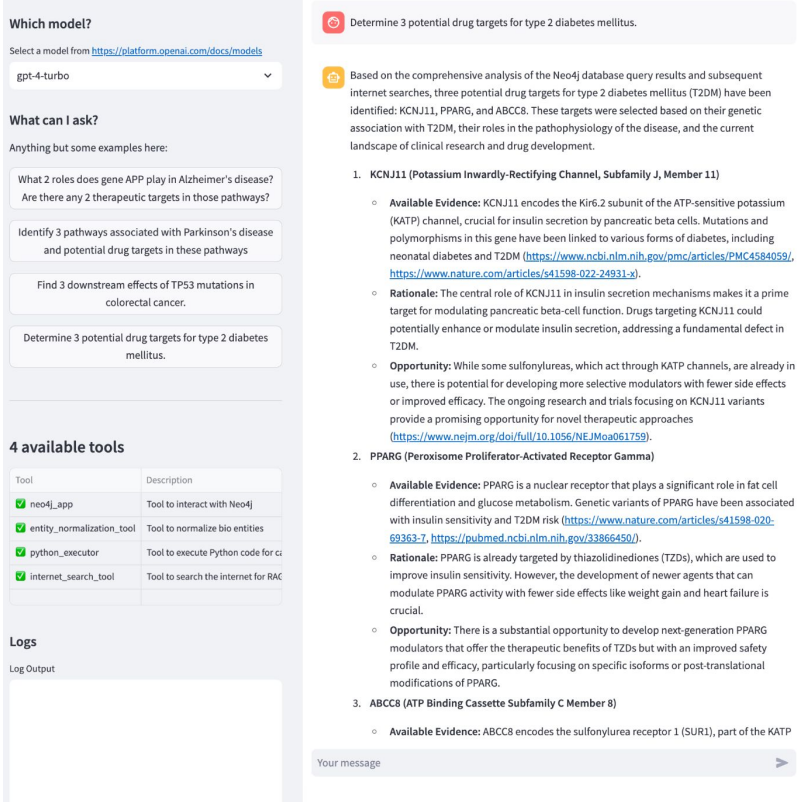
- I developed a ChemCrow-like agent from https://github.com/ur-whitelab/chemcrow-public
- Models used are from https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4
- Neo4J database built with:
  - Nodes
    - GO Terms: From Gene Ontology (GO) data.
    - Miscellaneous Entities: From KEGG like compounds and ‘map’ diseases.
    - Genes: From the HGNC dataset.
    - Pathways: From pathway datasets (e.g., KEGG, Reactome).
    - Diseases: From MONDO Disease Ontology.
  - Relationships
    - is_a: GO hierarchical relationships.
    - part_of: Gene-pathway associations.
    - has_molfunction, in_component, has_bioprocess: Gene-GO term interactions.
    - Has_Genetic_Association_With, Is_Differentially_Expressed_In, Affects_Pathway_In: Gene-disease associations from Open Targets.
    - Various interaction types: From KEGG data, such as "activates", "inhibits", etc.
- Other tools:
  - Internet Search
  - Python interpreter
  - Normalisation tool using API calls to mygene.info and mydisease.info
- Some prompts to set the tone and the reproducibility

# The Neo4J database



# Results for `3x T2DM targets`

## Screenshot of UI + response

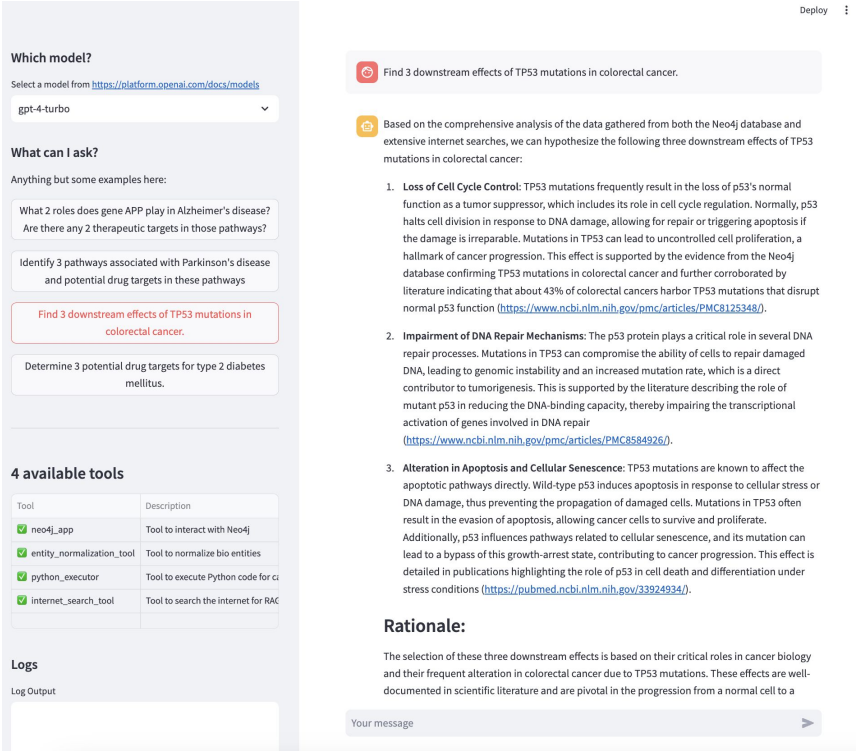



## My comments

- All 3 are well known, validated targets for a crowded disease like T2DM:
- ABCC8 is targeted by approved sulfonylureas (Gliclazide, Glipizide, Tolbutamide, Chlorpropamide, Gliquidone) and meglitinides (Nateglinide, Repaglinide), and others like Mitiglinide [source]
- PPARG is targeted by approved thiazolidinediones (Rosiglitazone, Pioglitazone) and investigational drugs (Troglitazone, Rivoglitazar, Naveglitazar…) [source]
- KCNJ11 is targeted by approved drugs such as Glimepiride, Diazoxide, Verapamil, and Glyburide, among others [source].
- **Pros**: the agent is able to find the well known targets using Neo4J + Internet queries without normalising entities using regex (see ./logs/history).
- **Cons**: For this assessment, I focused on well-known targets to highlight the limitations of LLMs when they primarily "recover" existing data using Retrieval-Augmented Generation (RAG) on documents or databases. This approach restricts hypothesis generation and creativity. It is the same with KEGG, once it is in KEGG it is well known and established in the biomedical domain. Limited use for new ideas. Results are limited to save time, money and limit the model output tokens and my assessment.

# Results for `TP53 effects in colorectal cancer`

## Screenshot of UI + response

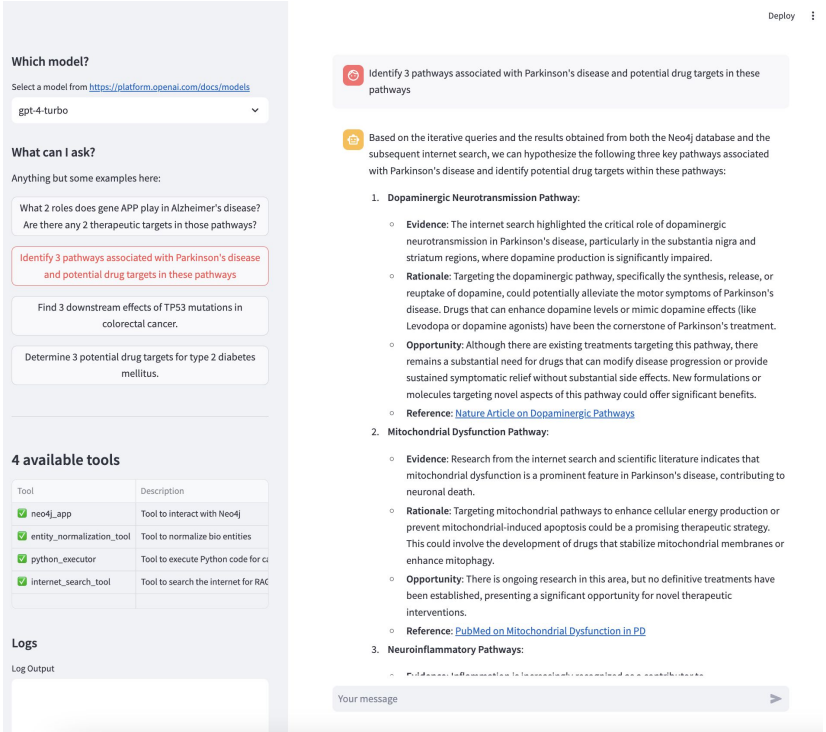



## My comments

- Lacks novelty and specificity: Outlines well-known TP53 mutation effects without new hypotheses specific to colorectal cancer (Hassin and Oren, 2023).
- General effects; lacks details: Missing detailed molecular pathways discussed in Hassin and Oren, 2023. But so does the underlaying database KEGG here.
- Cell cycle control loss: Correct but simplistic. Include pathways like CDK inhibitors (p21) relevant to colorectal cancer.
- DNA repair impairment: Accurate but vague. Should discuss pathways like homologous recombination affected by TP53 mutations, as detailed in the review.
- Apoptosis and senescence alteration: Correct but lacks context. Explore impacts on apoptotic pathways (BCL2 proteins) in colorectal cancer, as per the review.
- For the Rationale: Repeats known facts, lacks colorectal cancer specifics highlighted in the review.
- Opportunity: Generic therapeutic strategies. Recommend discussing recent colorectal cancer-specific advances from the review.
- **Summary**: Good foundation but needs depth, specificity, and novelty. Include detailed mechanisms, recent discoveries from full-text articles, and more pathways for impact, as elaborated in Hassin and Oren, 2023. Data from publications needs to be orchestrated into a database for this.

# Results for `targets in Parkinson pathways`

## Screenshot of UI + response




## My comments

- Dopaminergic Neurotransmission Pathway: Correct but basic. Enhance discussion on novel targets beyond dopamine levels and effects (e.g., targeting synaptic plasticity).
- Mitochondrial Dysfunction Pathway: Accurate but lacks detail. Discuss specific mitochondrial targets (e.g., PGC-1a activators, mitophagy enhancers) like Zong 2024.
- Neuroinflammatory Pathways: Correct but needs specificity. Suggest specific cytokines or microglial inhibitors being investigated (Tansey 2022)
- Rationale: Common knowledge in Parkinson's disease; lacks novel insights (Farrow 2022)
- Opportunity: Dopaminergic treatments well-established; focus more on novel mitochondrial and neuroinflammatory interventions for innovation
- Conclusion: Identified pathways and targets but need more detailed mechanisms and novel insights for significant impact.

# Future work

- Dedicated document database:
  - Better queries, with full text snippets, more precision and recall
  - Similar to TrendyGenes and PWAS (my papers)
- LangChain:
  - Recent Python library with constant updates. Unstable, difficult to build systems around it as it keeps changing
  - Not used due to problems with 1. async operations and 2. Latency problems
  - Not sure if I'll add it based on my experience, better to code yourself!
- More tools and data sources:
  - Target tractability: small molecules? Antibodies? Direction of the modulation?
  - Experimental data for novel targets with some tables
  - Essential: Competitor landscape for other therapies and pharma companies
- Visualisation tools for the Neo4J and histograms in the UI
- Other NER tools for rapidly querying
- Many more… target ID is a complicated multiparameter optimization problem!

# Other relevant efforts

- "Tx-LLM: A Large Language Model for Therapeutics" from Google
  - https://arxiv.org/pdf/2406.06316
- "LLM-Orchestrated Workflow Engine (LOWE) of Recursion data and tools"
  - https://www.recursion.com/lowe
- Enhanced Chat from EXAI…
- Many more…