

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

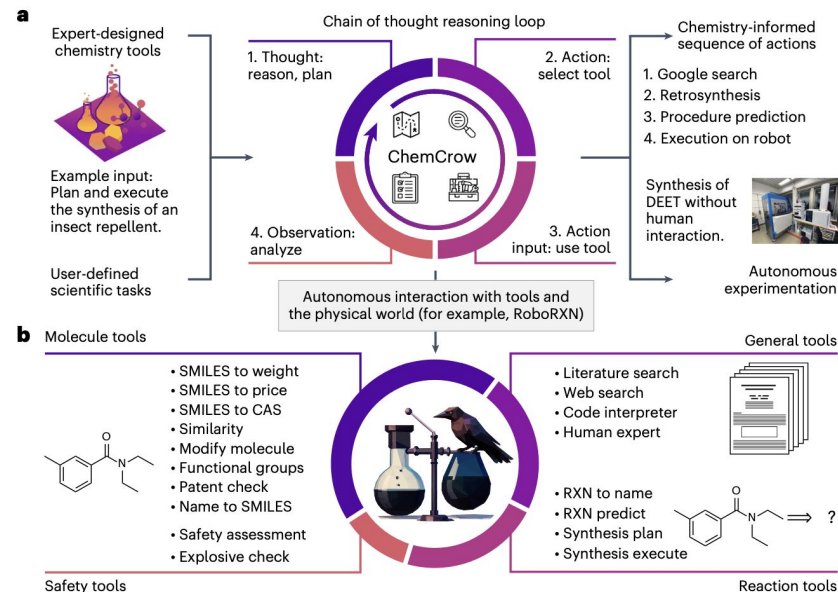
Accepted: 27 March 2024

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 Baldassari⁵,
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

**Database Information**

Use database

neo4j 

Node labels

GO (97,687)

Compound

Disease

GO

Gene

Map

Pathway

Relationship types

*(4,746,023)

Affects_Pathway_In

Has_Animal_Model_Evidence_For

Has_Genetic_Association_With

Has_Known_Drug_For

Has_Somatic_Mutation_In

Is_Differentially_Expressed_In

Is_Reported_In_Literature_For

activates_compound

activates_protein

binds_to_protein

dephosphorylates_protein

expresses_gene

has_bioprocess

has_molfunction

in_component

indirectly_affects

inhibits_compound

inhibits_protein

interacts_with_compound

is_a

missing_interaction

part_of

phosphorylates_protein

Property keys

0

Pathway

id

locus_group

name

score

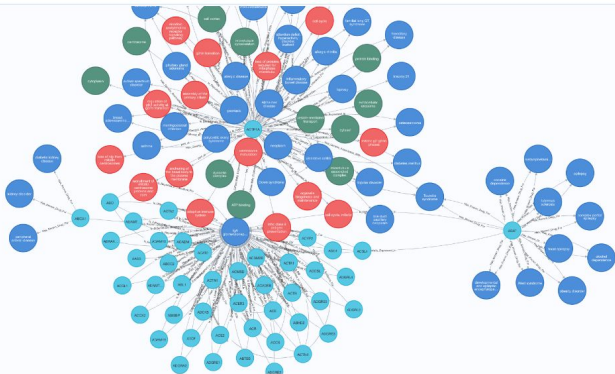
source

status

Connected as

neo4j\$

neo4j\$ MATCH p=()-[r:Has_Known_Drug_For]->() RETURN p LIMIT 25



Overview

Node labels

*(90)

Gene (45)

Disease (46)

Pathway (15)

GO (11)

Relationship types

*(220)

Has_Known_Drug_For (27)

Is_Differentially_Expressed_In (46)

Has_Animal_Model_Evidence_For (7)

Is_Reported_In_Literature_For (74)

Has_Genetic_Association_With (36)

part_of (15)

has_molfunction (2)

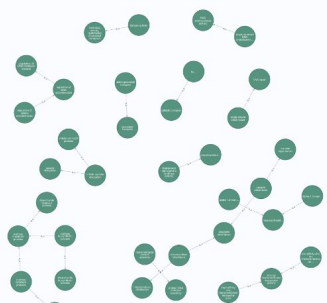
in_component (8)

Affects_Pathway_In (4)

has_bioprocess (1)

neo4j\$

neo4j\$ MATCH p=()->() RETURN p LIMIT 25



Overview

Node labels

*(39)

GO (39)

Relationship types

*(25)

is_a (25)

Displaying 39 nodes, 25 relationships.

Results for `3x T2DM targets`

Screenshot of UI + response

Which model?

Select a model from <https://platform.openai.com/docs/models>

gpt-4-turbo

What can I ask?

Anything but some examples here:

What 2 roles does gene APP play in Alzheimer's disease? Are there any 2 therapeutic targets in those pathways?

Identify 3 pathways associated with Parkinson's disease and potential drug targets in these pathways

Find 3 downstream effects of TP53 mutations in colorectal cancer.


Determine 3 potential drug targets for type 2 diabetes mellitus.


4 available tools

Tool	Description
<input checked="" type="checkbox"/> neo4j_app	Tool to interact with Neo4j
<input checked="" type="checkbox"/> entity_normalization_tool	Tool to normalize bio entities
<input checked="" type="checkbox"/> python_executor	Tool to execute Python code for cli
<input checked="" type="checkbox"/> internet_search_tool	Tool to search the internet for RAG

Logs

Log Output

 Determine 3 potential drug targets for type 2 diabetes mellitus.

 Based on the comprehensive analysis of the Neo4j database query results and subsequent internet searches, three potential drug targets for type 2 diabetes mellitus (T2DM) have been identified: KCNJ11, PPARG, and ABCC8. These targets were selected based on their genetic association with T2DM, their roles in the pathophysiology of the disease, and the current landscape of clinical research and drug development.

- KCNJ11 (Potassium Inwardly-Rectifying Channel, Subfamily J, Member 11)**
 - Available Evidence:** KCNJ11 encodes the Kir6.2 subunit of the ATP-sensitive potassium (KATP) channel, crucial for insulin secretion by pancreatic beta cells. Mutations and polymorphisms in this gene have been linked to various forms of diabetes, including neonatal diabetes and T2DM (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4584059/>, <https://www.nature.com/articles/s41598-022-24931-0>).
 - Rationale:** The central role of KCNJ11 in insulin secretion mechanisms makes it a prime target for modulating pancreatic beta-cell function. Drugs targeting KCNJ11 could potentially enhance or modulate insulin secretion, addressing a fundamental defect in T2DM.
 - Opportunity:** While some sulfonylureas, which act through KATP channels, are already in use, there is potential for developing more selective modulators with fewer side effects or improved efficacy. The ongoing research and trials focusing on KCNJ11 variants provide a promising opportunity for novel therapeutic approaches (<https://www.nejm.org/doi/full/10.1056/NEJMoa061759>).
- PPARG (Peroxisome Proliferator-Activated Receptor Gamma)**
 - Available Evidence:** PPARG is a nuclear receptor that plays a significant role in fat cell differentiation and glucose metabolism. Genetic variants of PPARG have been associated with insulin sensitivity and T2DM risk (<https://www.nature.com/articles/s41598-020-69363-7>, <https://pubmed.ncbi.nlm.nih.gov/33866459/>).
 - Rationale:** PPARG is already targeted by thiazolidinediones (TZDs), which are used to improve insulin sensitivity. However, the development of newer agents that can modulate PPARG activity with fewer side effects like weight gain and heart failure is crucial.
 - Opportunity:** There is a substantial opportunity to develop next-generation PPARG modulators that offer the therapeutic benefits of TZDs but with an improved safety profile and efficacy, particularly focusing on specific isoforms or post-translational modifications of PPARG.
- ABCC8 (ATP Binding Cassette Subfamily C Member 8)**
 - Available Evidence:** ABCC8 encodes the sulfonylurea receptor 1 (SUR1), part of the KATP

Your message

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, Glipizide) and meglitinides (Nateglinide, Repaglinide), and others like Mitiglinide [[source](#)]
- PPARG is targeted by approved thiazolidinediones (Rosiglitazone, Pioglitazone) and investigational drugs (Trogliptazone, 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

Which model?

Select a model from <https://platform.openai.com/docs/models>

gpt-4-turbo

What can I ask?

Anything but some examples here:

What 2 roles does gene APP play in Alzheimer's disease?
Are there any 2 therapeutic targets in those pathways?

Identify 3 pathways associated with Parkinson's disease
and potential drug targets in these pathways

Find 3 downstream effects of TP53 mutations in
colorectal cancer.

Determine 3 potential drug targets for type 2 diabetes
mellitus.

4 available tools

Tool	Description
<input checked="" type="checkbox"/> neo4j_app	Tool to interact with Neo4j
<input checked="" type="checkbox"/> entity_normalization_tool	Tool to normalize bio entities
<input checked="" type="checkbox"/> python_executor	Tool to execute Python code for ci
<input checked="" type="checkbox"/> internet_search_tool	Tool to search the internet for RAG

Logs

Log Output

Deploy

Find 3 downstream effects of TP53 mutations in colorectal cancer.

Based on the comprehensive analysis of the data gathered from both the Neo4j database and extensive internet searches, we can hypothesize the following three downstream effects of TP53 mutations in colorectal cancer:

1. **Loss of Cell Cycle Control:** TP53 mutations frequently result in the loss of p53's normal function as a tumor suppressor, which includes its role in cell cycle regulation. Normally, p53 halts cell division in response to DNA damage, allowing for repair or triggering apoptosis if the damage is irreparable. Mutations in TP53 can lead to uncontrolled cell proliferation, a hallmark of cancer progression. This effect is supported by the evidence from the Neo4j database confirming TP53 mutations in colorectal cancer and further corroborated by literature indicating that about 43% of colorectal cancers harbor TP53 mutations that disrupt normal p53 function (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8125348/>).

2. **Impairment of DNA Repair Mechanisms:** The p53 protein plays a critical role in several DNA repair processes. Mutations in TP53 can compromise the ability of cells to repair damaged DNA, leading to genomic instability and an increased mutation rate, which is a direct contributor to tumorigenesis. This is supported by the literature describing the role of mutant p53 in reducing the DNA-binding capacity, thereby impairing the transcriptional activation of genes involved in DNA repair (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8584926/>).

3. **Alteration in Apoptosis and Cellular Senescence:** TP53 mutations are known to affect the apoptotic pathways directly. Wild-type p53 induces apoptosis in response to cellular stress or DNA damage, thus preventing the propagation of damaged cells. Mutations in TP53 often result in the evasion of apoptosis, allowing cancer cells to survive and proliferate. Additionally, p53 influences pathways related to cellular senescence, and its mutation can lead to a bypass of this growth-arrest state, contributing to cancer progression. This effect is detailed in publications highlighting the role of p53 in cell death and differentiation under stress conditions (<https://pubmed.ncbi.nlm.nih.gov/33924934/>).

Rationale:

The selection of these three downstream effects is based on their critical roles in cancer biology and their frequent alteration in colorectal cancer due to TP53 mutations. These effects are well-documented in scientific literature and are pivotal in the progression from a normal cell to a

Your message

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 underlying 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

Which model?

Select a model from <https://platform.openai.com/docs/models>

gpt-4-turbo

What can I ask?

Anything but some examples here:

What 2 roles does gene APP play in Alzheimer's disease?
Are there any 2 therapeutic targets in those pathways?

Identify 3 pathways associated with Parkinson's disease
and potential drug targets in these pathways

Find 3 downstream effects of TP53 mutations in
colorectal cancer.

Determine 3 potential drug targets for type 2 diabetes
mellitus.

4 available tools

Tool	Description
✓ neo4j_app	Tool to interact with Neo4j
✓ entity_normalization_tool	Tool to normalize bio entities
✓ python_executor	Tool to execute Python code for ci
✓ internet_search_tool	Tool to search the internet for RAC

Logs

Log Output

Deploy ⓘ

Identify 3 pathways associated with Parkinson's disease and potential drug targets in these pathways

Based on the iterative queries and the results obtained from both the Neo4j database and the subsequent internet search, we can hypothesize the following three key pathways associated with Parkinson's disease and identify potential drug targets within these pathways:

- Dopaminergic Neurotransmission Pathway:**
 - Evidence:** The internet search highlighted the critical role of dopaminergic neurotransmission in Parkinson's disease, particularly in the substantia nigra and striatum regions, where dopamine production is significantly impaired.
 - Rationale:** Targeting the dopaminergic pathway, specifically the synthesis, release, or reuptake of dopamine, could potentially alleviate the motor symptoms of Parkinson's disease. Drugs that can enhance dopamine levels or mimic dopamine effects (like Levodopa or dopamine agonists) have been the cornerstone of Parkinson's treatment.
 - Opportunity:** Although there are existing treatments targeting this pathway, there remains a substantial need for drugs that can modify disease progression or provide sustained symptomatic relief without substantial side effects. New formulations or molecules targeting novel aspects of this pathway could offer significant benefits.
 - Reference:** [Nature Article on Dopaminergic Pathways](#)
- Mitochondrial Dysfunction Pathway:**
 - Evidence:** Research from the internet search and scientific literature indicates that mitochondrial dysfunction is a prominent feature in Parkinson's disease, contributing to neuronal death.
 - Rationale:** Targeting mitochondrial pathways to enhance cellular energy production or prevent mitochondrial-induced apoptosis could be a promising therapeutic strategy. This could involve the development of drugs that stabilize mitochondrial membranes or enhance mitophagy.
 - Opportunity:** There is ongoing research in this area, but no definitive treatments have been established, presenting a significant opportunity for novel therapeutic interventions.
 - Reference:** [PubMed on Mitochondrial Dysfunction in PD](#)
- Neuroinflammatory Pathways:**
 - Evidence:** Research from the internet search and scientific literature indicates that neuroinflammation is a prominent feature in Parkinson's disease, contributing to neuronal death.
 - Rationale:** Targeting neuroinflammatory pathways to reduce inflammation or enhance neuroprotection could be a promising therapeutic strategy. This could involve the development of drugs that inhibit pro-inflammatory cytokines or enhance anti-inflammatory responses.
 - Opportunity:** There is ongoing research in this area, but no definitive treatments have been established, presenting a significant opportunity for novel therapeutic interventions.
 - Reference:** [PubMed on Neuroinflammation in PD](#)

Your message

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-1α 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...