Drug Discovery using Generative AI

ITSOLERA GEN_AI **TEAM BETA**

Overview

Project Overview: AI Powered Drug Discovery

This project is designed to systematically retrieve and analyze data related to approve drugs from the ChEMBL database. ChEMBL is a highly respected chemical database that provides bioactive molecule information, making it a valuable resource for drug discovery and research. The primary focus of this project is to extract key information, such as the date of drug approval and the corresponding SMILES (Simplified Molecular Input Line Entry System) notation, which serves as a textual representation of the chemical structures. This project aims to facilitate deeper insights into the drug discovery through fine tuning of LLM on retrieval data of ChEMBL.

Key Features:

- **Comprehensive Data Retrieval:** The project employs advanced data retrieval techniques to efficiently gather a wide array of data from the ChEMBL database, specifically targeting drugs that have achieved full approval status. This includes critical identifiers and chemical structure information that are foundational for research.
- **SMILES Notation Extraction:** SMILES notation is a compact and widely used chemical notation. The project extracts and stores this information for each approved drug, enabling easy analysis and integration into molecular modeling tools.
- **Approval Timeline Analysis:** By capturing the approval dates of each drug, the project allows for the creation of temporal insights, helping to understand trends and shifts in drug approvals over time.
- **Fine Tuning of LLM (ChatCohere):** The retrieval data is stored in CSV file and use to fine tune the Cohere LLM. The LLM provides the preferred name of drug, detailed information of the target drug, and the structure of the drug.

- **Interactive Web Interface (gardio):** Provides user_friendly web interface whereuser/chemist can easily enter the ChEMBL ID and SMILES of the drug and receives drug target information along with drug structure as well as preferred name of the drug.

Technical Components

- **Libraries:** Datamol, Chembl_warehouse_client, Pandas, ChatCohere
- **APIs:** Cohere
- **Web Framework:** Gardio

Workflow

- **User Input:** ChEBL ID and SMILES of a drug is provided through gardio interface,
- **Drug Target Information:** Cohere LLM is queried to provide drug target information. This information includes a detailed summary of the drug's known biological target and its mechanism of action. It also provides approaches towards the synthesis of new compounds approaching novel drugs for multi targets, give chemical and physical properties of the drug as well. Moreover, it provides target protein structures and formulas for the drug.
- **Output:** Drug target information, preferred name of the drug, and molecular structure of the drug are presented to user

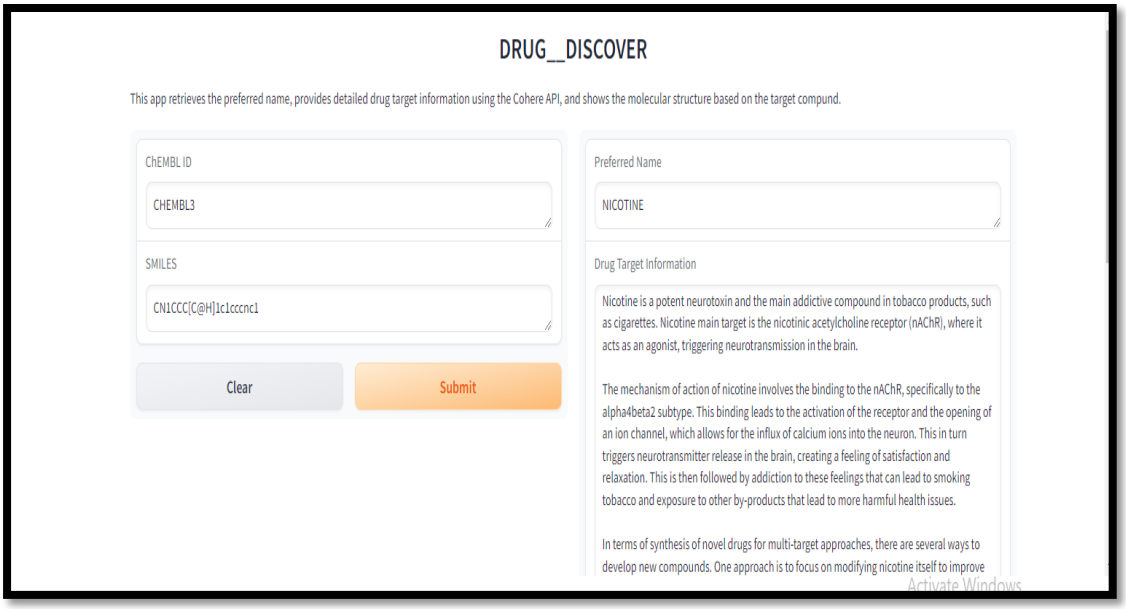

Figure 1: Gardio based web interface showing user input and output

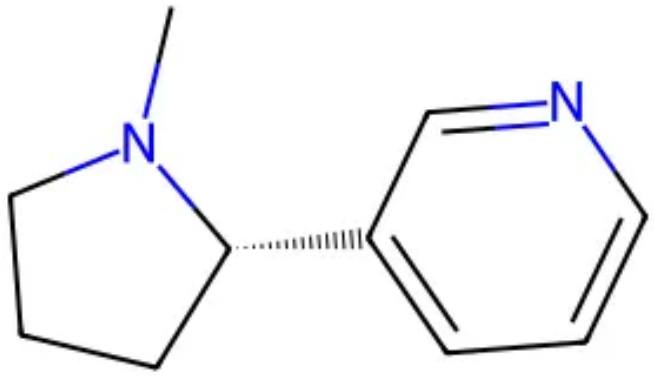

Figure 2: Molecular structure of the targeted drug

Potential Applications

- **Targeted Drug Recommendations:** The fine-tuned Cohere LLM can analyze a patient's specific biological markers or disease characteristics and recommend drugs that are most likely to be effective. This enables more precise and personalized treatment plans, improving patient outcomes.
- **Precision Oncology:** In cancer treatment, the model could identify drugs that target specific mutations or cancer subtypes, offering tailored therapy options that align with the molecular profile of the patient's tumor.

Key Innovations

Integration of Data Retrieval with Language Model Fine-Tuning:

- **Seamless Data Pipeline:** The project innovatively integrates a robust data retrieval process from the ChEMBL database with the fine-tuning of a Cohere LLM. This seamless pipeline enables the extracted data to be directly utilized in enhancing the language model's capabilities, ensuring that the insights generated are highly relevant and up-to-date.
- **End-to-End Automation:** By automating the entire workflow from data retrieval to insight generation, the project significantly reduces the manual effort required in traditional drug research, making the process more efficient and scalable.

Fine-Tuning of LLM for Specialized Drug Information:

- **Domain-Specific Knowledge Embedding:** The fine-tuned Cohere LLM is specialized in the pharmaceutical domain, embedding extensive knowledge about drug structures, target interactions, and approval histories. This allows the model to generate highly accurate and context-specific information, which is crucial for advanced drug research.
- **Enhanced Predictive Capabilities:** The LLM's ability to provide detailed drug information and predict drug-target interactions is a significant leap forward. This innovation allows for more precise predictions in virtual screening and drug repurposing, reducing the time and cost associated with experimental validation.