Chemical Recommender System (CRS)

The chemical recommender system has been created to programmatically suggest similar and less toxic replacements for targeted small molecules. For questions, contact panair@sandia.gov. An overview of the program structure exists in the following steps.

1. Structural Similarity Shortlisting

Similarity shortlisting is done using RDkit to calculate molecular fingerprints for the query and all candidates from the previous PubChem CIDs. The fingerprints are implemented as 2048-bit vectors that are set on/off to describe the existence of structural properties. Using the Tanimoto similarity formula, we calculate structural similarity of all candidates against the query.

This process is streamlined with Milvus, a modern vector databasing solution, by preprocessing and indexing the vectors to complete the search in a matter of minutes. At this point a variety of user specified filters can be applied such as substructure matching or excluding heavy metals. The candidates are resorted, and the best are shortlisted for further searching.

2. OPERA Comparisons

Many properties of the chemicals are predicted using an open-source software called OPERA. OPERA utilizes a weighted k-nearest neighbor approach using a minimum number of descriptors calculated using PaDEL. Each of the following operations are computed on both the query and the shortlist of candidates. Then, comparisons are calculated by assigning a comparison score to each candidate for Structural similarity, Thermophysical similarity, and Toxicity. Opera is used for:

- Structural Similarity: Molecular Weight, Number of Rings, Number of Lipinski Failures
- Thermophysical Similarity: Melting Point, Boiling Point, Log P, Vapor Pressure, Henry's Law constant
- Toxicity: Log BCF, CATMoS EPA, CATMoS LD50

3. Synthetic Accessibility Scoring

The RDkit package provides a value from 1(easy to synthesize) to 10 (hard to synthesize) for molecules based off smiles. The SA Scoring is factored into the CRS recommendations.

4. Data Preparation

The comparison metrics for are distinctly normalized and aggregated for a final similarity score. Users are given the option to adjust the weightages of each of these values in computation, allowing for reranking based on preferred similarity types.

4. Availability/Developer

The CRS is accessible as a web app or through a CLI, both of which automatically generate extensive graphs and a pdf report on the search results. Both are available to users from online container services, so the entire program is available by running a single docker-compose file. The CLI version gives developers additional options in using the CRS to add their own models. If a user has created their own ML models to integrate with the CRS, it can be wrapped in a docker container and be easily provided through CLI argument.